Juking the Stats: Policing, Misreporting, and Policy Evaluation[*]

Michael Gibilisco[†] Carlo M. Horz[‡]

January 2024

Abstract

We analyze a game-theoretic model of crime and crime reporting to study the quality of crime statistics. A citizen potentially engages in illicit behavior; an enforcement agency chooses effort and how to report outcomes. Because of signaling concerns, the agency may misreport. We show that multiple equilibria can exist and characterize when crime is under- or over-reported. Increasing the agency's costs of data manipulation can backfire, increasing misreporting in crime statistics. When calculating treatment effects of parameter changes, the effect on reported statistics will not equal the effect on true statistics, and the true effect can be under- or overestimated.

[*] First draft: May 2023. Thanks to Emiel Awad, Casey Crisman-Cox, Germán Gieczewski, Alex Hirsch, Federica Izzo, Brenton Kenkel, Moritz Marbach, Joe Ruggiero, and Miguel Rueda for comments and discussions. This project has benefited from audience feedback at the Caltech Applied Political Theory Lab, Emory's Theory & Models Working Group, and the Southern California Political Institutions and Political Economy Conference.

[†] Assistant Professor, Division of Humanities & Social Sciences, California Institute of Technology. Email: `michael.gibilisco@caltech.edu`.

[‡] Assistant Professor, Department of Political Science, Texas A&M University, College Station, TX 77845-4348 Email: `carlo.horz@tamu.edu`.

1 Introduction

The maintenance of public order is a key task for all polities and a precondition for economic development and peaceful relations between citizens (Olson 1993). For these reasons, societies frequently feature specialized agencies that are tasked with enforcing laws. Most prominently, this is the police, but in the United States for example, other law enforcement agencies include the Drug Enforcement Administration, Customs and Border Protection, or the Bureau of Alcohol, Tobacco, Firearms and Explosives. Besides maintaining public order, they are also often asked to keep accurate records of illicit behavior and enforcement outcomes. Thus, these kinds of agencies have a dual role in creating data: they behave in a certain way, thus affecting outcomes, and they record these outcomes for their own and other actors' use. For example, police departments enforce local laws, e.g., issue speeding tickets or patrol neighborhoods, and keep records such as crime statistics.

These law enforcement statistics play a key role in the public discourse and in scholarly work. For example, whether crime increased or decreased can be an important determinant for electoral choices and outcomes (Arnold and Carnes 2012). Scholars also use crime statistics to investigate the extent to which there is racial bias in police stops (Fryer Jr 2019; Knox, Lowe and Mummolo 2020) or to quantify the effects of policing (Blair et al. 2021; Levitt 2002), sanctions (Bell, Jaitman and Machin 2014; Kovandzic, Vieraitis and Boots 2009), or gun control (Dube, Dube and García-Ponce 2013; Duggan 2001). Similarly, statistics published by immigration agencies play a role in analyzing if restrictive migration policies are effective or not (Czaika and Hobolth 2016).

Both scholars and practitioners worry about the dual role of enforcement agencies in data creation. Agents in charge of creating records may be tempted to tamper with reports in order to avoid accountability for improper behavior or to increase the chances of influencing political decisions, such as their agency's future funding or voters' electoral choices. While systematic evidence of manipulated crime statistics is, by their nature, difficult to gather, anecdotal evidence suggests that manipulated crime statistics can be a salient problem. One notable example is documented by whistle-blower Adrian Schoolcraft, a former officer in the New York City Police Department. When his superiors remained unresponsive to his concerns about the accuracy of crime statistics, he provided audio recordings to the press that revealed the extent to which police officers were under intense pressure to find ways to lower reported crime (Reuters 2012).[1] Notably, officers were told to

> "refuse certain robbery reports in order to manipulate and lower official crime statistics so that the neighborhood appeared safer. Command precinct person-

[1] This seems to be a persistent problem in New York City. For example, in 2004, in the midst of bargaining over higher wages, police union leaders argued that the city's crime statistics were falsified by local police officers (New York Times 2004). Similar stories appear in other major cities (British Broadcasting Corporation 2013; Dallas Morning News 2020; Los Angeles Times 2015).

> nel recount calling crime victims to discourage them from making complaints. Officers were encouraged to downgrade felony thefts to petty larcenies or misdemeanors. Officers were instructed to convert robbery reports to the category of lost property" (Eterno and Silverman 2017, 4).

Given the prominence of law enforcement statistics in both scholarly work and public discourse, it is important to understand to what extent such misreporting incentives affect the quality of data and inferences. Unfortunately, social scientists currently lack a theoretical framework that explicitly models both the *actual* incidence of crime and the *reported* incidence of crime. Developing such a framework faces two important challenges. First, the actual incidence of crime is tainted by strategic interaction between law enforcement officers and citizens pondering whether or not to engage in illicit behavior. Hence, actual law enforcement statistics are driven by both expected and actual illicit behavior and enforcement effort. Second, if law enforcement officers have signaling concerns when submitting their reports, i.e., they anticipate that crime reports allow other actors to learn about their behavior or characteristics, then the benefits of misreporting are not exogenous. The reason is that the credibility of reported crime depends on third-party actors' beliefs about actual crime incidence. This creates a link between actual and reported crime and a dependency in officers' enforcement and reporting behavior.

Taking into account these challenges, we provide a game-theoretic model of law enforcement behavior and reporting. Consistent with the dual role of enforcement agencies, the model consists of two stages: an enforcement stage that represents an encounter between an agent and a potential citizen target, and a subsequent reporting stage in which the agent can potentially misrepresent the enforcement outcome. We characterize equilibria of the game and use the characterization to derive substantive implications about measurement error in crime statistics and bias when estimating causal effects on crime outcomes.

Our game proceeds as follows: in the enforcement stage, the target and the agent play an inspection game in which the target chooses whether or not engage in illicit behavior, and the agent chooses to exert high or low enforcement effort. Behavior in the enforcement stage determines a binary law enforcement statistic, i.e., either a crime was committed or not. For much of the analysis, we assume that illicit behavior by the target increases the likelihood of the statistic indicating a crime whereas agent effort decreases the likelihood of the statistic indicating a crime.[2] The agent observes the true crime statistic; she subsequently decides how to record it in the reporting stage. Crime can be reported even if it did not occur ("over-reporting"), or no crime can be reported even if it occurred ("under-reporting"). The target faces some (opportunity) costs of engaging in illicit behavior but may be tempted to do so if he anticipates low agent effort. The agent wishes to prevent crime, but exerting effort

[2] We also consider an alternative version of the model in which, all else equal, enforcement effort increases the likelihood of the statistic indicating a crime, e.g., speeding tickets.

is costly. Moreover, the agent also wishes to signal that she exerted effort. Specifically, the agent internalizes the posterior probability of exerting effort *conditional on the reported crime statistic.* The assumption is consistent with the presence of a third-party actor, such as a relevant community or politician, who observes the reported crime statistic, and decides on a level of support or funding.[3] The agent cares about support, and the optimal support level increases in the posterior assessment of the agent exerting effort. Finally, tampering with data is costly (e.g., the agent faces the possibility of audits), so whenever the agent does not report the true crime statistic, she incurs some costs.

In the game's equilibrium, the enforcement and reporting stages are intertwined. When the agent reports, she is tempted to lie when doing so means that she increases the posterior probability of high effort. If policing effort decreases the crime statistic, then conditional on a crime occurring, the agent is tempted to cover it up by reporting that no crime occurred. If policing effort increases the crime statistic (as with speeding tickets for example), then conditional on no crime occurring, the agent is tempted to inflate crime numbers by over-reporting. In either case, the exact incentive to lie depends on the difference in posteriors after the two potential reports, which depends on the equilibrium probability that the agent exerts effort. Conversely, the reporting stage also affects the enforcement stage. The agent anticipates that her effort choice influences the crime statistic and hence the opportunity (or necessity) to manipulate data when reporting. Furthermore, we show that there exist complementarities between incentives for enforcement effort and incentives to misreport in equilibrium.

Besides elucidating the incentives for enforcement effort with endogenous reporting, we find three key results. First, despite the fact that, separately, the reporting (sub)game and the enforcement game have a unique equilibrium, taken together, the game can have multiple equilibria. The reason is that the reporting stage influences the enforcement stage to such an extent that expectations can be self-fulfilling. When high effort is expected, exerting high effort can be profitable for the agent because a subsequent report of no crime will signal high effort. When low effort is anticipated, in contrast, exerting high effort may not be profitable because a report of no crime will signal either misreporting or little criminal activity. Such a finding has two important implications. First, even with identical characteristics, law enforcement organizations can differ widely in terms of both the accuracy of reports as well as their enforcement behavior because equilibrium expectations determine enforcement and reporting outcomes. Our result therefore resonates with empirical work emphasizing the importance of "culture," i.e., the importance of leadership and managerial expectations, for the workings of police departments (Cordner 2017; Ingram, Terrill and Paoline 2018; Johnson 2015; Terrill, Paoline and Manning 2003). Second, theories of policing often model deterrence using inspection games which generally have unique predictions

[3] We provide a more extensive discussion of this assumption below.

about equilibrium behavior (Avenhaus, Von Stengel and Zamir 2002). In contrast, our inspection game *with endogenous reporting* can have multiple equilibria and more complicated comparative statics even when the underlying inspection and reporting games are simple. Misreporting can therefore complicate efforts that use equilibrium properties to estimate the preferences of police (as in Antonovics and Knight 2009; Knowles, Persico and Todd 2001; Stashko 2022, for example).

Second, in contrast to existing intuitions (e.g., Cook and Fortunato 2022), we show that an increase in the agent's data manipulation costs can both increase or decrease data quality. We consider two substantively different notions of data quality: the ex-ante probability of misclassification and the difference between the probability of a true crime and the probability of a reported crime. We show that in our framework, these notions coincide, and data quality is the probability that the enforcement stage produces the outcome in which the agent lies multiplied by the probability with which the agent actually misreports given this outcome. Thus, measurement error is a function of both behavior in the enforcement game (illicit activity and agent effort) and misreporting. An increase in data manipulation costs has a direct effect of decreasing misreporting, but the agent also increases her level of effort. This decreases the frequency with which the state in which the agent lies occurs (which further reduces measurement error), but increases the credibility of the agent's report. As a result, the agent manipulates more often, and this indirect effect can offset the other effects. Hence, a policy intervention aimed at increasing data quality by making it more costly to tamper with data can backfire.

Third, we characterize the bias when computing causal effects, and demonstrate that it can be positive or negative. Specifically, our model allows us to compute treatment effects on both true and observed crime statistics. We demonstrate that the treatment effect of a parameter on observed crime is an additive function of the treatment effect on actual crime and a bias term. Importantly, the bias is equal to (the negative of) the effect of the parameter on equilibrium measurement error. Hence, if a variable increases or decreases measurement error either directly or indirectly, the bias term will not be zero. This result is relevant for any study that uses administrative law enforcement statistics in order to assess the causal effect of various political and economic variables on crime.[4]

To illustrate it, we apply this result to the voluminous literature that investigates how opportunity costs affect criminal behavior as measured by crime statistics—see Khanna et al. (2021) and Bell, Bindler and Machin (2018) for examples and Draca and Machin (2015) for a review. Consistent with the general result, the observed treatment effect can be higher or lower than the true treatment effect. The reason is that the treatment

[4] For example, Di Salvatore (2019) analyzes the effect of United Nations peacekeepers on crime. Charnysh (2019) shows that migrant diversity has a positive effect on crime. Jassal (2020), Magaloni, Franco-Vivanco and Melo (2020), and Blair, Karim and Morse (2019) show that newly formed police units with a particular composition can affect crime rates. Finally, Dynes and Holbein (2020) demonstrate that the partisanship of a government has only a limited effect on crime outcomes.

changes measurement error relative to the control. When it decreases measurement error, the observed effect overestimates the true effect because it captures a decrease in under-reporting, hence more reported crime. When it increases measurement error, the observed effect underestimates the true effect because it captures an increase in under-reporting, i.e., less reported crime.

Our framework also has two broader implications. First, administrative data may not always be of higher quality compared with data derived from media reports. Scholars are well aware that media reports of crime or violence (or other kinds of illicit behavior) are often argued to be skewed by selection effects; for example, more successful illicit behavior is less likely to be reported. Administrative data are often thought to be a remedy to this problem (Berman, Shapiro and Felter 2011; Horz and Marbach 2022; Shaver et al. 2022; Weidmann 2016). However, our analysis shows that the strategic incentives by reputation-seeking agents may make administrative data problematic proxies of reality as well. The key insight is that internal politicking, careerism, and signaling can lead to strategically misreported statistics even when members of the agency do not expect the data to be released to the wider public. Although our approach is theoretical, this finding is also echoed in Garbiras-Díaz and Slough (2022) who empirically document the misreporting of Colombian bureaucrats using a novel audit study.

Second, we show that a special, especially pernicious case studied by the methodological literature on measurement bias—misclassification that is *not* conditionally random—is the generic case for law enforcement data. In this context, conditional randomness means that a treatment does not influence the probability with which the variable is misclassified (Bound, Brown and Mathiowetz 2001; Hausman, Abrevaya and Scott-Morton 1998; Meyer and Mittag 2017; Weidmann 2016). In our model, this probability corresponds to an "interim" notion of measurement error—it is the equilibrium probability that the agents misreports conditional on an enforcement outcome. Because the agent optimally conditions their misreporting decision on expected effort and expected crime, if a treatment has an causal effect on the true crime statistic, then it must have an effect on the equilibrium misclassification probability, and hence directly influence the observed crime statistic. Thus, generically, conditional randomness cannot hold for data created by enforcement agencies.

These results contribute to several literatures. First, there is a relatively recent theoretical literature concerned with police learning. McCall (2019) and Hübert and Little (2023) examine models of policy experimentation within police agencies and show how inferences and accompanying decisions affect policing disparities across demographic groups. The former contribution focuses on how different police tactics encourage or discourage help from residents while the latter contribution focuses on imperfect learning—some police officers may not accurately condition on relevant variables. By contrast, we focus on the ability of social scientists and third-parties to learn from potentially misreported policing data and how police officers or agencies are affected by their own, anticipated (mis-)reporting.

Second, there is a theoretical literature on endogenous data quality and collection outside of the crime context. Gibilisco and Steinberg (2022) study collateral damage data quality in conflict settings and compare measurement error in government and NGO reports. In contrast to our paper, the costs of misreporting are endogenized through a potential audit, but the outcome that can be misreported is an exogenous state of the world. Alonso and Câmara (2023) study how organizations' data governance policies (i.e., data tampering prevention and detection) incentivize agents to produce more or less useful data. In their model, an agent designs an experiment and subsequently reports an outcome to a principal who has to decide whether or not to adopt a project advocated by the agent. They identify a tradeoff between informative experiments and truthful reporting. By contrast, we do not study how principals optimally design rules to gather useful data. Rather, our data is endogenously generated by the agent's effort choice and a citizen's choice to engage in illicit activity. Moreover, we focus on the quality of crime data as a function of background factors (e.g., the agent's costs to engage in data manipulation and the citizen's level of economic opportunity) and the associated inferences that can be drawn from misreported data. Because the agent's effort choice is an intrinsic outcome of interest—it prevents crime—our paper is also related to Roger (2013) who studies the classic moral hazard model in which an agent can misreport an outcome. Roger (2013) then examines the implications for optimal contracting by the principal. Finally, similar to us, Patty and Penn (2015) are interested in understanding data quality and measurement from a formal theory perspective. However, they employ an axiomatic approach, linking empirical measures with certain properties.[5]

Third, there is an empirical literature that focuses on enforcement agency behavior. A small set directly tackles strategic misreporting: Luh (2022) exploits a change in data reporting practices in Texas to study racial bias through systematically misreported trooper reports. Eckhouse (2022) focuses on rape and how performance management may have contributed to police officers reclassifying instances of rape as "unfounded." Cook and Fortunato (2022) examine how state legislative capacity enhances the transparency and quality of statistics reported by local police agencies. Relatedly, Arora (2023) shows that juvenile crime is under-recorded relative to adult crime. Another set of papers looks for evidence of racial profile by examining policing tactics and enforcement outcomes data with a close connection to theoretic models of policing (Antonovics and Knight 2009; Anwar and Fang 2006; Clark et al. 2020; Knowles, Persico and Todd 2001; Stashko 2022). Our model builds on these contributions in designing the enforcement stage but enriches them by

[5] More broadly, a recent line of work studies empirical research designs through theoretical models (Bueno De Mesquita and Tyson 2020; Slough 2023; Slough and Tyson 2022). Similar to these papers, we identify a friction in causal inference. Different from these papers, we focus on strategic misreporting as a distinct challenge for accurate measurement and causal inference.

explicitly introducing a reporting stage. Finally, although their focus is on the relationship between policing and crime, Ba et al. (2021) combine a matching model and crime data to quantify the effects of police assignment mechanisms to local neighbors on crime rates.

2 Model

There are two players: an agent A, and a potential target T. To aid our presentation, we use "he" to refer to T and "she" to refer to A. The game has three pieces: first, the agent and target interact in an enforcement game; second, the enforcement game produces a law enforcement outcome; third, the agent chooses how to record the statistic. We discuss each in turn.

For the enforcement game, the agent chooses effort $e \in \{0,1\}$, where $e = 1$ means high effort and $e = 0$ means low effort. The potential target chooses $c \in \{0,1\}$, where $c = 1$ means engaging in illicit behavior and $c = 0$ means not engaging in illicit behavior. An outcome of the enforcement game is a pair (e, c). The agent and target have preferences over these outcomes. We write the target's payoffs as

$$u_T^{\texttt{en}} = c(1-e) - \gamma c.$$

Thus, the target receives normalized benefit of 1 for illicit activity when there is low enforcement effort. The parameter γ represents the (opportunity) costs of illicit activity and is private information to the target. In particular, it is drawn from an absolutely continuous random variable with convex support, cumulative distribution function (CDF) G, and an associated probability density function (PDF) g. We assume that g is continuous on its support, ensuring that derivatives of implicitly defined quantities are continuous.

For the agent's payoffs, we write

$$u_A^{\texttt{en}} = \beta ec - \rho e.$$

Here, $\beta > 0$ is the intrinsic benefit of exerting high effort with illicit activity, e.g., motivations to catch the target engaging in the act. This specification is isomorphic to $-\beta(1-e)c$ replacing βec, in which case β represents the cost of letting criminal activity happen under the agent's watch when she exerts low effort. The agent's relative cost of high effort is ρ and is private information. It is drawn from an absolutely continuous random variable with convex support, CDF F, and PDF f, where f is continuous over its support.

The timing for the enforcement game is as follows. First, the private costs ρ and γ are realized for A and T, respectively. Subsequently, the players choose their actions simultaneously.

The enforcement interaction is standard (e.g., Clark et al. 2020; Knowles, Persico and

Todd 2001; Luh 2022; Stashko 2022).[6] Previous contributions usually abstract away from explicitly discussing how this interaction maps into law enforcement statistics, however. Given that data manipulation is our focus, we now make assumptions on how outcomes of the enforcement interaction map onto (reported) law enforcement statistics. We think of the law enforcement statistic a partial summary of the outcomes of the enforcement interaction. What they summarize depends on the interpretation of enforcement effort, the interpretation of illicit activity, the possibility of third-party reporting , and the dimensionality of the law enforcement statistic. Of course, the law enforcement statistic may not be sole outcome that the agent and target care about.

Specifically, let $x(e, c) \in \{0, 1\}$ denote the true law enforcement outcome. We call $x = 1$ the crime outcome (or law violation) and $x = 0$ no crime outcome (no violation of the law). We mainly consider the following *idealized* specification for the data generating process:

$$x = (1 - e)c.$$

This specification reflects a focus on preventative enforcement effort, which decreases the opportunities for criminal behavior. In other words, the outcome of enforcement interaction consists of the opportunities for crime. The agent's choice of effort, $e = 1$, shrinks opportunities (e.g., patrolling) while planning behavior by the target, $c = 1$, expands opportunities (e.g., arming). We assume that opportunities for crime exists if and only if $(1 - e)c = 1$.

The specification encapsulates several assumptions. First, the target uses any opportunities for crime, so crime occurs if and only if $(1 - e)c = 1$. The agent will learn of crime passively. Moreover, planning $c = 1$ is not detectable or not a major crime. For example, the target is a gang deciding whether or not to engage in a turf war. If the police patrol, then the target may see this patrol, and subsequently, the target may simply run away. In turn, the outcome $x = 1$ represents gang violence, which is only possible if the target arms and the agent does not patrol. Here, β is relative cost of letting crime occur on agent's watch, and γ is the opportunity cost of arming for turf war. Naturally, the crime statistic outcome $x = 1$ implies that no effort was exerted, $e = 0$.

After the enforcement stage, the agent observes the outcome x and her data manipulation costs, which are denoted by η. The agent writes a report $\tilde{x} \in \{0, 1\}$. We interpret $\tilde{x} = x$ as a truthful report and $\tilde{x} \neq x$ as a lie, which refer to misclassifying a statistic or misreporting a statistic. Consistent with the definition of the true law enforcement statistic x, we call $\tilde{x} = 1$ reported crime and $\tilde{x} = 0$ reported no crime.

We now discuss the total payoffs for the agent, which are the sum of payoffs from the

[6] Similar interactions appear in the terrorism and counterterrorism literature (e.g., Di Lonardo and Dragu 2021; Dragu 2011)

enforcement and reporting stages:

$$u_A = u_A^{\mathtt{en}} + u_A^{\mathtt{re}} = \underbrace{\beta ec - \rho e}_{\text{enforcement payoff}} + \underbrace{b_{\tilde{x}} - \eta \mathbb{I}[\tilde{x} \neq x]}_{\text{reporting payoff}}.$$

Here, $b_{\tilde{x}} \equiv \Pr(e = 1 \mid \tilde{x})$ is the *endogenous* posterior belief of high enforcement effort given a report $\tilde{x}$. Thus, our agent wants to convince a third party that it exerted high enforcement effort. Such a third party could be a funding source like the mayor or a city council, a manager higher up in the chain of command like the chief of police, or the citizenry at large where the agent has an easier job when citizens believe she is working hard. The agent's cost of data manipulation is η. It is a random variable drawn from a CDF H with PDF h. As above, H has convex support, and h is continuous over that support. We also assume that η is bounded below by 0 so $\min \operatorname{supp}(H) = \underline{\eta} \geq 0$. Finally, the parameter η is realized only after enforcement game.

To summarize, the full sequence of the model is the following.

1. A observes the cost of effort $\rho \sim F$, and T observes the opportunity cost $\gamma \sim G$.
2. Simultaneously, A chooses effort e and T chooses behavior c.
3. Enforcement payoffs are realized, $u_i^{\mathtt{en}}$ for $i = T, A$.
4. The law enforcement statistic is produced according to $x = (1 - e)c$.
5. A observes the realization of the statistic x and cost of manipulating data $\eta \sim H$.
6. A writes a report $\tilde{x} \in \{0, 1\}$.
7. A receives reporting payoffs $u_A^{\mathtt{re}} = b_{\tilde{x}} - \eta \mathbb{I}[\tilde{x} \neq x]$.

2.1 Definition of equilibrium and quantities of interest

For the target, a strategy is a function $s_T : \operatorname{supp}(G) \to \{0, 1\}$, and $s_T(\gamma) = 1$ is the decision to engage in illicit activity given costs γ. For the agent, a strategy is a tuple $s_A = (s_A^{\mathtt{en}}, s_A^{\mathtt{re}})$, where $s_A^{\mathtt{en}} : \operatorname{supp}(F) \to \{0, 1\}$ and $s_A^{\mathtt{re}} : \{0, 1\} \times \operatorname{supp}(H) \to \{0, 1\}$. Here, $s_A^{\mathtt{en}}(\rho)$ is the effort decision given cost ρ, and $s_A^{\mathtt{re}}(x, \eta)$ is the report given the true law enforcement statistic $x \in \{0, 1\}$ and manipulation costs η.[7] Recall beliefs are $b_{\tilde{x}} = \Pr(e = 1 \mid \tilde{x})$ for reports $\tilde{x} \in \{0, 1\}$. We focus on perfect Bayesian equilibria referred to as equilibria hereafter. Specifically, an equilibrium is an assessment (s, b) where (i) $s = (s_T, s_A)$ is a sequentially rational strategy profile given beliefs $b = (b_0, b_1)$ and (ii) beliefs b are consistent with the strategies and updated via Bayes' rule whenever possible.

It is often useful to work with the following higher-order choice probabilities. Given

[7] Note that the agent's reporting strategy $s_A^{\mathtt{re}}$ does not depend on e, which is without loss of generality as A's reporting payoffs do not depend on e.

strategy profile s, define the following:

$$\begin{aligned} C^s &= \Pr(c = 1 \mid s) = \int \mathbb{I}[s_T(\gamma) = 1] g(\gamma) d\gamma, \\ E^s &= \Pr(e = 1 \mid s) = \int \mathbb{I}[s_A^{\texttt{en}}(\rho) = 1] f(\rho) d\rho, \text{ and} \\ R_x^s &= \Pr(\tilde{x} \neq x \mid x, s) = \int \mathbb{I}[s_A^{\texttt{re}}(x, \eta) \neq x] h(\eta) d\eta. \end{aligned}$$

In words, C^s is probability of illicit activity given strategy profile s, E^s the probability of high effort, and R_x^s the probability of misreporting outcome x.

Besides examining the forces that shape equilibrium behavior, we are interested in two quantities that, in equilibrium, are influenced by the players' behavior and that play important roles in empirical analyses. Our first definition is about measurement error. Recall that for a discrete random variable, measurement error *is* misclassification. Therefore, we define measurement error as follows.

Definition 1. *Given a strategy profile s, measurement error M^s is defined as the ex-ante probability of misclassification:*

$$M^s \equiv \Pr(x \neq \tilde{x} \mid s).$$

.

We interpret M^s as a summary measure of data quality. As Definition 1 emphasizes, it is computed without knowing true or reported crime.[8] Hence, it can also be thought as *average* data quality because $M^s = \Pr(x = 1|s) R_1^s + \Pr(x = 0|s) R_0^s$. We are interested in how the quantity M^s responds to changes in parameters. In particular, we will investigate whether ex-ante misclassification decreases if it becomes more costly to the agent to manipulate data.

Another quantity of interest is the difference between the probabilities of actual and recorded crime, formally defined in the next definition.

Definition 2. *Given a strategy profile s, the difference in the probability of a crime and the probability of a reported crime is*

$$D^s \equiv \Pr(x = 1|s) - \Pr(\tilde{x} = 1 \mid s).$$

For brevity, we sometimes refer to D^s as difference in crime probabilities or difference in crime coverage. The quantity is important because it can be intuitively linked to over- and under-reporting. In particular, if $D^s < 0$, then crime is over-reported; if $D^s > 0$, then

[8] After the game has concluded, and the random variables x and $\tilde{x}$ are realized, there is (ex-post) misclassification if $x \neq \tilde{x}$. After the enforcement stage, after x is realized, (interim) misclassification occurs with probability $\Pr(x \neq \tilde{x} \mid x)$ denoted by R_x^s with strategy profile s. We return to the connection between interim and ex-ante misclassification below.

it is under-reported. Furthermore, given that most descriptive and inferential statistics use means, the difference in actual and reported crime is of particular substantive importance.[9] While Definitions 1 and 2 appear different, we subsequently show that the quantities are closely linked in equilibrium, i.e., up to their sign, they are identical. Thus, insights for one are directly relevant to the other.

In Section 4.3, we use the model to study to what extent causal effect of parameters may differ for actual and observed crime. We defer our definition of causal effects until we have characterized the equilibrium, however.[10]

2.2 Discussion of assumptions

We make several simplifying assumptions. First, in order to incorporate both an enforcement and a reporting stage, both stages are deliberately stylized. In particular, the enforcement stage features a single choice by the agent and the target, even though in reality it consists of a sequence of choices. For example, in Knox, Lowe and Mummolo (2020) and Clark et al. (2020), the police officer can both stop a target and, conditional on stopping, choose a certain tactic. Moreover, the utility functions of the agent and target are deliberately sparse (with as few parameters as possible) so that the target's benefit of a successful crime is normalized to 1 and there is no separate parameter for being caught and punished, although such a parameter is easily accommodated. Finally, the agent intrinsically cares about fighting criminal behavior with a parameter β. This assumption has empirical support (Stashko 2022), but it would be isomorphic if the agent cares about crime occurring on her watch.

In addition, the baseline model assumes that enforcement effort suppresses the crime statistic x. This means we are primarily studying preventative enforcement. Below, we also briefly consider a version of the model with remedial enforcement, i.e., $x = ec$.[11] Here, the outcome of enforcement interaction is the uncovering criminal behavior. The agent's choice of effort, $e = 1$, is searching for a crime (stopping) while the target's choice, $c = 1$, is the execution of a crime (speeding). We show that whether crime is under- or over-reported in our framework depends on the nature of enforcement effort—see Proposition 5 below.

Relatedly, in our crime production technology, the agent can powerfully affect crime: enforcement effort renders crime impossible. Although this may be plausible for well-staffed, well-resourced police agencies and for some types of crime, the assumption may be too strong for other applications. In Appendix D, we generalize the crime production technology

[9] Definition 1 may be more substantively appealing quantity for measurement error, however, because it generalizes to variables that have more than two values.

[10] Briefly, we define causal effects of a parameter θ on outcomes x and $\tilde{x}$ in terms of comparative statics. Hence, in contrast to Definition 1 and 2, the definition relies on an equilibrium characterization, and not just on a strategy profile.

[11] By using the terminology preventative and remedial enforcement, we are borrowing the terminology in Dragu and Przeworski (2019).

to allow for less than perfect crime preventation after high effort, assuming that $\Pr(x = 1|e, c) = \alpha(1-e)c + (1-\alpha)c$. Here, the strength of crime prevention is parametrized by $\alpha \in [0, 1]$. Intuition may suggest that for the kinds of crimes in which α is smaller, misreporting is less of an issue because crime outcomes contain less information about the agent's effort. As a result, data quality should be higher and measurement error should be lower. In Appendix D, we characterize the equilibria in this more general model and show that this intuition is not always correct. Indeed, we demonstrate that for some parameter values, decreasing α (making crime less informative about agent effort) can increase measurement error.[12] The intuition is that a lower level of α *discourages* the agent from exerting effort, which makes the crime outcome $x = 1$ more likely. As discussed below, measurement error increases as the probability of crime increases, all else equal.

Especially important is our choice of how to model signaling concerns in the agent's utility function. The agent is assumed to care about the perception of effort given the reported crime statistic $\tilde{x}$. On the one hand, this is consistent with previous work that directly embeds beliefs into an agent's utility function to capture signaling or reputational concerns (as in Fox and Van Weelden 2012; Kartik and Van Weelden 2019). These concerns might be critical to a retention, financing, or public support decision by a principal. For example, police agencies may want to convince mayors, city councils, state legislators, police chiefs, or citizens that they worked hard to lower crime. On the other hand, this is a departure from standard moral hazard models where the signaling or reputational concerns are related to a type, e.g., competence or alignment of policy interests.[13]

To aid in interpretation, we show that our baseline model is a close approximation to the case in which agents have a type, and they internalize the posterior probability of being a "diligent" rather than "lazy" type—see Appendix C.[14] This information structure is more complicated, however, because it involves an additional round of updating, leading to more complicated posterior beliefs.[15] In order to focus on the potential misreporting of crime statistics given endogenous behavior in an enforcement encounter, we focus on the simpler case in which the agent is assumed to directly care about perceived effort, treating the case where there are agent types as a robustness exercise.[16]

[12] In the limit $\alpha = 0$, there is no misreporting and measurement error is zero.

[13] See Fearon (1999) who shows that in accountability models, when the voter cares about effort provided by the incumbent, voter-optimal equilibria break down when there is small heterogeneity across politicians.

[14] The key assumptions are that (i) the lazy type has effort costs $\rho = \infty$ whereas the diligent type has costs ρ drawn from F, and (ii) that both types (diligent and lazy) face the same distribution of data manipulation costs. Appendix C contains the details.

[15] Given that effort is still unobserved, they still must form beliefs about the effort choice, and use it to form a posterior about the agent being the good type.

[16] Another possible setup is that the agent has private information about the "severity" of the crime problem, i.e., to what extent reducing crime is a difficult task, and is tempted to over-report crime to signal that the situation is dire to obtain additional funding. This can be accommodated by our framework if we were to assume that the agent internalizes third-party beliefs about the severity of a problem. Similar to the current setup, here, misreporting and effort will be intertwined and a model is needed to study how misreporting affects inference. To begin studying this issue, we focus on the more tractable incentives to

Finally, recall that η is the cost of data manipulation. The cost is not an inherent feature of the agent but rather situational to the enforcement outcome. More specifically, the exact manipulation costs are only known after the enforcement stage. They depend on unmodeled details of the case.[17] For example, burglaries are generally not violent crimes but they would be if they involve an occupied residence and a weapon (i.e., the difference between burglary and attempted robbery). Thus, a criminal incident might be classified as violent depending on whether or not a weapon is involved, but the presence of a weapon might be easier or harder to misclassify depending on specific conditions. If the weapon is a firearm for example, then police may have a harder time reporting no weapon was present when the firearm was discharge than when it was not. Finally, one can also interpret η as an outcome of a future audit. To see this, suppose that the punishment for a false report is constant $\omega > 0$, and q that is the probability of getting caught in a lie, with q being a random variable drawn from distribution with support on $[0, 1]$. Then define $\eta \equiv q\omega$, where q depends on facts of the case. With this interpretation, a key assumption is that the audit happens only after the benefit from the third-party's belief is realized. For example, good performance might be rewarded with more overtime or better choice of scheduling or vehicles.

3 Equilibrium Characterization

We focus on a particular class of equilibria, what we call full-support equilibria.

Definition 3. *An equilibrium* (s, b) *has full support if* $E^s \in (0, 1)$*,* $C^s > 0$*,* $R_0^s = 0$ *and* $R_1^s \in (0, 1)$*.*

In words, Definition 3 says that a full-support equilibrium (s, b) will entail uncertainty about whether or not the agent exerts high effort and whether or not the true law enforcement statistic is a crime outcome (because $0 < (1 - E^s)C^s < 1$). Furthermore, in a full-support equilibrium, the law enforcement statistic is truthfully reported after the no-crime outcome ($x = 0$), but will be reclassified with some probability strictly between zero and one after the crime outcome ($x = 1$).

Our focus on full-support equilibria is motivated by the fact that it is likely that both law enforcement outcomes are reported in the data, i.e., the data generating process produces no-crime ($\tilde{x} = 0$) and crime ($\tilde{x} = 1$) reports with positive probability. In this case, the next result says that the equilibrium satisfies certain properties. Appendix A contains the proof.

Lemma 1. *If* (s, b) *is an equilibrium such that both reports are sent, i.e.,* $\Pr(\tilde{x} = 1 \mid s) \in (0, 1)$*, then the following hold:*

signal high effort.

[17] Below, we explicitly analyze the consequences of situational characteristics such as body cameras in our treatment of changes to data manipulation costs.

1. *the no-crime outcome is never misreported, and the crime outcome is not always misreported (i.e., $R_0^s = 0$ and $R_1^s < 1$);*
2. *illicit activity can occur (i.e., $C^s > 0$); and*
3. *the agent is guaranteed not to always exert effort (i.e., $E^s < 1$).*

An implication of Lemma 1 is that if both reports are sent in equilibrium, the only potential for misreporting results from the enforcement outcome, $x = 1$. Furthermore, comparing Lemma 1 and Definition 3 shows that focusing on equilibria in which both reports are sent produces most of the substantive properties of full-support equilibria, except for two conditions. First, a positive probability of effort, $E^s > 0$, guarantees posterior beliefs $b_{\tilde{x}}$ are not trivial. Second, a positive probability of misreporting after a crime outcome, $R_1^s > 0$, guarantees misreporting occurs with positive probability. These two additional properties are empirically plausible as well: at least some of the time, law enforcement officers exert effort and, given widespread concerns about the validity of law enforcement statistics, may manipulate their reports. Finally, we provide sufficient conditions that guarantee all equilibria have full support in Proposition B.1 in Appendix B.[18]

To characterize full-support equilibria, recall that we focus on preventative policing where $x = (1-e)c$. Consider the posterior beliefs after each report, $b_{\tilde{x}}$. When $R_0^s = 0$, a reported crime, $\tilde{x} = 1$ implies that there indeed was a crime, $x = 1$, which also implies no effort, $e = 0$. Hence, $\Pr(e = 1|\tilde{x} = 1, s) = b_1 = 0$.

After observing $\tilde{x} = 0$, i.e., reported no crime, the agent may have misreported the crime statistic, so that the posterior probability of the agent having exerted effort is:

$$\begin{aligned}\Pr(e = 1|\tilde{x} = 0, s) = b_0 &= \frac{\Pr(\tilde{x} = 0|e = 1, s)\Pr(e = 1|s)}{\Pr(\tilde{x} = 0|s)} \\ &= \frac{E^s}{E^s + (1 - E^s)[(1 - C^s) + C^s R_1^s]}.\end{aligned}$$

To understand this expression, note that the posterior can be written in terms of prior beliefs and the informativeness of the report $\tilde{x}$. Hence, if $E^s > 0$, we can rewrite b_0 as

$$b_0 = \left(1 + \frac{1 - E^s}{E^s}\frac{1 - C^s + C^s R_1^s}{1}\right)^{-1}.$$

The term $\frac{1-E^s}{E^s}$ is the prior ratio and $\frac{1-C^s+C^sR_1^s}{1}$ is the likelihood ratio. The latter is equal to $\frac{\Pr(\tilde{x}=0|e=0,s)}{\Pr(\tilde{x}=0|e=1,s)}$, i.e., how likely the signal $\tilde{x} = 0$ is if the agent exerted effort (denominator) or not (numerator).

It is now straightforward to derive the agent's equilibrium reporting strategy. Given that $x = 1$ and realized misreporting cost η, her net-of-enforcement payoff is $b_1 = 0$ after

[18] The conditions are standard. They hold, e.g., if F and G have full support over the real line and H is the uniform over $[0, \bar{\eta}]$ with $\bar{\eta} \geq 1$.

reporting truthfully but $b_0 - \eta$ after misreporting. Thus, a threshold strategy is optimal: the agent lies after $x = 1$ if and only if $\eta < b_0 - b_1$. Define the difference in posterior beliefs as the *manipulation stakes*:

$$\bar{\mu}(E^s, C^s, R_1) \equiv b_0 - b_1.$$

Note that $\bar{\mu}$ increases in E^s and C^s, but decreases in R_1^s. The equilibrium threshold for manipulation, $\hat{\eta}^*$, solves the following condition:

$$\bar{\mu}\left(E^s, C^s, H(\hat{\eta})\right) = \hat{\eta}. \tag{1}$$

Hence, the equilibrium misreporting rate is $R_1^s = H(\hat{\eta}^*)$. Note that the agent does not condition on effort, which is irrelevant in the reporting stage, but beliefs about effort enter her strategy implicitly. The reason is that in order to form accurate beliefs, any third-party observer takes into account their prior belief about effort E^s and illicit behavior C^s. In particular, more expected effort or criminal behavior makes $\tilde{x} = 0$ more credible, increasing the agent's misreporting incentives.

Lemma 2. *The agent manipulates more if more effort or illicit activity is expected:*

$$\frac{\partial \hat{\eta}^*}{\partial E^s} = -\frac{\frac{\partial \bar{\mu}}{\partial E^s}}{\frac{\partial \bar{\mu}}{\partial R_1^s} h(\hat{\eta}^*) - 1} > 0 \qquad \textit{and} \qquad \frac{\partial \hat{\eta}^*}{\partial C^s} = -\frac{\frac{\partial \bar{\mu}}{\partial C^s}}{\frac{\partial \bar{\mu}}{\partial R_1^s} h(\hat{\eta}^*) - 1} > 0.$$

Now consider the enforcement game. Given a realized opportunity cost γ, the target's expected utility for engaging in illicit behavior is $1 - E^s - \gamma$. By contrast, the expected utility of not engaging in illicit behavior is 0. Hence, he chooses to engage in illicit behavior if and only if $\gamma < \hat{\gamma}^*$ where

$$\hat{\gamma}^* = 1 - E^s. \tag{2}$$

Thus, $C^s = G(\hat{\gamma}^*)$. Note that the target's best response is decreasing in E^s: the more likely the agent exerts effort, the more likely illicit activity is unsuccessful, and hence the lower the incentives are to commit it.

For the agent, if she exerts effort, then at the enforcement stage, she catches illicit activity in the act with probability C^s, obtaining β. In addition, the agent has to pay the costs of effort, ρ. Moreover, because there will be no crime ($x = 0$) and the agent correctly reports this ($\tilde{x} = 0$), the agent obtains a payoff of b_0 at the reporting stage. Thus, the expected payoff of exerting effort is

$$\underbrace{\beta C^s - \rho}_{\substack{\text{enforcement}\\ \text{payoff}}} + \underbrace{b_0}_{\substack{\text{reporting}\\ \text{payoff}}}.$$

However, if the agent shirks, the enforcement payoff is 0. For the reporting stage, with

with probability $1 - C^s$, there will be no actual crime, $x = 0$. Since the agent accurately reports $\tilde{x} = 0$, the agent's reporting payoff is again b_0. However, with probability $C^s(1-R_1^s)$, there is a crime but the agent's data manipulation costs are relatively high, so that no misreporting takes place. The agent's payoff in this case is $b_1 = 0$. With probability $C^s R_1^s$, there is an actual crime but the data manipulation costs are small, and so the agent misreports. The agent's payoff is $b_0 - \mathbb{E}[\eta | \eta \leq \hat{\eta}^*]$. Thus, the expected payoff for not exerting effort is:

$$(1 - C^s)b_0 + C^s \left[(1 - R_1^s)0 + R_1^s \left(b_0 - \mathbb{E}[\eta|\eta \leq \hat{\eta}^*]\right)\right].$$

Comparing these two expressions, the agent works if and only if

$$\rho < C^s(\beta + \Psi(E^s, C^s)),$$

where

$$\begin{aligned}\Psi(E^s, C^s) &= \underbrace{(1 - R_1^s)\bar{\mu}(E^s, C^s, R_1^s)}_{\text{relative reward incentive}} + \underbrace{R_1^s \mathbb{E}[\eta|\eta \leq \bar{\mu}(E^s, C^s, R_1^s)]}_{\text{expected manipulation costs}} \\ &= (1 - H(\hat{\eta}^*))\hat{\eta}^* + \int_{\underline{\eta}}^{\hat{\eta}^*} \eta h(\eta) d\eta.\end{aligned}$$

In words, Ψ captures the agent's dynamic incentives to work, and these incentives are strictly increasing in the manipulation stakes, which in equilibrium are equal to the threshold $\hat{\eta}^*$. As such, Lemma 2 implies that Ψ is increasing in expected effort E^s and expected illicit activity C^s.

Thus the agent use a threshold strategy such that $E^s = F(\hat{\rho})$. In addition, target's best response to $\hat{\rho}$ is $C^s = G(1 - F(\hat{\rho}))$. Plugging this into the preceding expressions, an equilibrium (s, b) is characterized by a threshold strategy, $\hat{\rho}^*$, that solves

$$\underbrace{G(1 - F(\hat{\rho}))[\beta + \Psi(F(\hat{\rho}), G(1 - F(\hat{\rho})))]}_{\equiv \Lambda(\hat{\rho})} = \hat{\rho}. \tag{3}$$

Proposition 1 summarizes the analysis thus far.

Proposition 1. *If (s, b) is a full-support equilibrium, then the following hold:*

1. *The agent exerts effort if and only if $\rho < \hat{\rho}^*$ where ρ^* solves Equation 3, so $E^s = F(\hat{\rho}^*)$.*

2. *The target engages in illicit behavior if and only if $\gamma < \hat{\gamma}^*$ where $\hat{\gamma}^*$ solves Equation 2, so $C^s = G(\hat{\gamma}^*)$.*

3. *In the reporting subgame, the agent never misreports the no-crime outcome $x = 0$, but misreports after the crime outcome $x = 1$ if and only if $\eta < \hat{\eta}^*$, where $\hat{\eta}^*$ solves Equation 1. So $R_1^s = H(\hat{\eta}^*)$.*

To better understand equilibrium incentives, consider the relationship between enforcement and reporting as encapsulated by the term Ψ. It is clear that the agent's reporting strategy affects these dynamic incentives: if the equilibrium threshold for manipulation ($\hat{\eta}^*$) increases, Ψ increases as well, i.e., $\frac{\partial \Psi}{\partial \hat{\eta}^*} > 0$. The above analysis implies that effort increases.

The agent's equilibrium effort also affects reporting through beliefs $\bar{\mu}$. When more effort is expected, i.e., the equilibrium threshold $\hat{\rho}^*$ increases, two effects emerge. First, there is a partial effect given by $\frac{\partial \hat{\eta}^*}{\partial \hat{\rho}^*} = \frac{\partial \hat{\eta}^*}{\partial E^s} f(\hat{\rho}^*) > 0$. Substantively, as reports of no crime are more credible, the agent is convinced to lie more often. Second, there is a total effect:

$$\frac{d\hat{\eta}^*}{d\hat{\rho}^*} = \frac{f(\hat{\rho}^*)}{1 - \frac{\partial \overline{\mu}}{\partial R_1^s} h(\hat{\eta}^*)} \underbrace{\left(\frac{\partial \overline{\mu}}{\partial E^s} - \frac{\partial \overline{\mu}}{\partial C^s} g(1 - F(\hat{\rho}^*)) \right)}_{\text{Total, weighted effect on stakes}},$$

which accounts for, via the target's best response, the fact that higher effort decreases criminal activity. This total effect, weighted by the responsiveness of the target to higher effort, can be positive or negative, depending on parameter values. We show below that the sign of this term is crucial for equilibrium uniqueness and comparative statics.

3.1 Examining full-support equilibria

Complements or substitutes? It is useful to recast our analysis in terms of whether effort and lying are complements or substitutes. To do so, recall that the agent has two instruments at her disposal: providing high effort and manipulating data. If the agent provides effort in a full-support equilibrium, then the agent will not lie. The contrapositive is if the agent manipulated data, then the agent provided low effort. Thus, when looking at *realized actions*, the agent's instruments are substitutes. However, now consider the incentives to manipulate and the incentives to exert effort. The former is increasing in expected effort ($\frac{\partial \bar{\mu}}{\partial E^s} > 0$) and the latter is increasing in expected manipulation (because $\frac{\partial \Psi}{\partial \hat{\eta}} > 0$). Thus, when looking at *expected equilibrium rates*, the agent's instruments are complements.

Uniform manipulation costs While Lemma 2 implies that manipulation is increasing in effort and illicit activity , the precise relationship between these quantities can nevertheless be complex. To see this, consider the case in which the data manipulation costs are drawn uniformly over $[0, 1]$. Because $R_1 = H(\hat{\eta}) = \hat{\eta}$, Equation 1 simplifies to

$$\frac{E^s}{E^s + (1 - E^s)\left[1 - C^s + C^s \hat{\eta}\right]} = \hat{\eta}.$$

Hence, the equilibrium threshold can be explicitly computed as

$$\hat{\eta}^* = \frac{-\left[1 - C^s(1 - E^s)\right] + \sqrt{\left[E^s + (1 - E^s)(1 - C^s)\right]^2 + 4E^s(1 - E^s)C^s}}{2(1 - E^s)C^s}.$$

Besides the explicit solution for the equilibrium manipulation threshold $\hat{\eta}^*$, uniform manipulation costs also simplify the expression of Ψ. Because $H(\hat{\eta}) = \hat{\eta}$ and $\int_0^{\hat{\eta}} \eta d\eta = \frac{\hat{\eta}^2}{2}$, we can write Ψ as a function of the manipulation threshold:

$$\Psi(\hat{\eta}) = (1 - \hat{\eta})\hat{\eta} + \frac{\hat{\eta}^2}{2} = \hat{\eta} - \frac{\hat{\eta}^2}{2},$$

which is increasing in $\hat{\eta}$. Although Ψ is a relatively simple function of the manipulation threshold $\hat{\eta}$, it is a complicated function of E^s and C^s, because, even with uniform manipulation costs, the equilibrium manipulation threshold is a complicated function of both.

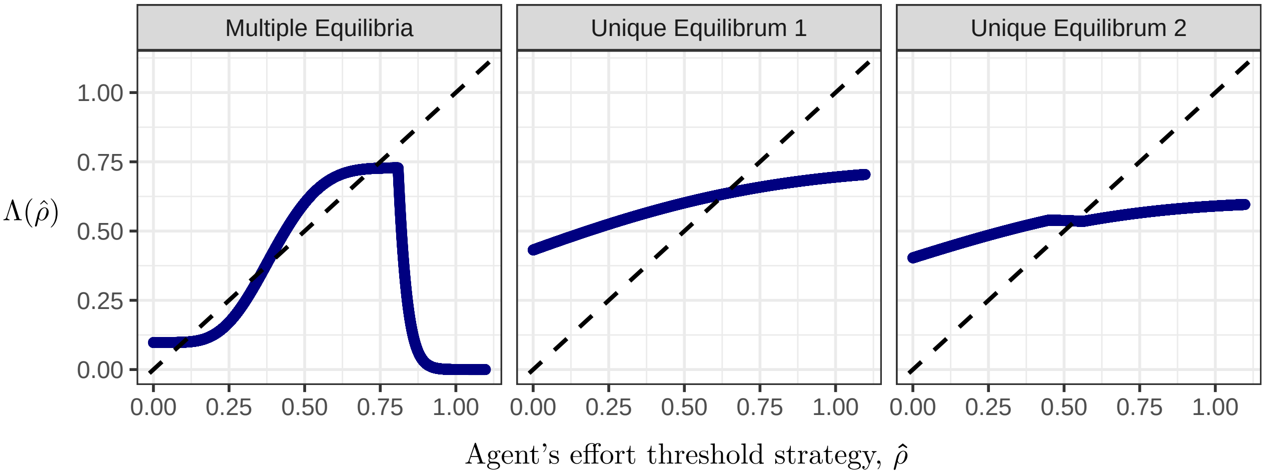
Equilibrium multiplicity Multiple solutions to Equation 3 can exist, so the model may admit multiple equilibria. This is perhaps surprising because if reporting were guaranteed to be truthful, i.e., $\tilde{x} = x$, then there would be a unique equilibrium in the enforcement game as in standard inspection games. Likewise, if behavior in the enforcement game were treated as exogenous, i.e., E^s and C^s are fixed constants, then there would a unique equilibrium in the reporting stage. When enforcement and reporting are endogenously determined, multiplicity can arise, however.

To see the intuition for this, consider the left panel in Figure 1. The horizontal axis represents possible values for the agent's enforcement threshold strategy, and the vertical axis graphs $\Lambda(\hat{\rho})$ from Equation 3. In this example, the agent's intrinsic incentives for enforcement effort are small as $\beta = 0.1$. Thus, the agent's incentives for providing effort are largely dynamic and given by $C^s\Psi$. Recall that the difference in posterior beliefs $\bar{\mu}$ describes the manipulation stakes, where larger stakes lead to larger dynamic incentives for effort, Ψ. Both terms are increasing in expected equilibrium effort E^s. If E^s is large, there are large stakes in the reporting game and large incentives for effort, but if E^s is small, there are small stakes and little incentives for effort.

Thus, expectations about enforcement effort can be self-fulfilling. When the third party believes the agent is likely to exert effort, there are large dynamic incentives to work. When the third party believes the agency is unlikely to exert effort, there are small dynamic incentives to work. This creates the multiple equilibrium in Figure 1's left panel. Technically, this multiplicity does not appear is traditional inspection games, which are the standard way to model deterrence in policing. Thus, the substantive features of equilibrium play in crime models can depend on the ability of police agents to misreport outcomes.

Substantively, multiple equilibria imply that several combinations of different report-

Figure 1: Examples of equilibrium multiplicity and uniqueness.



Notes: In the left panel, $\beta = 0.1$, $\eta \sim \mathcal{N}(0, 1.5)$, $\rho \sim \mathcal{N}(0.5, 0.1)$, and γ is drawn $\mathcal{U}(0, 1\mathrm{e}^{-3})$ with probability 0.95 and drawn $\mathcal{U}(0, 2)$ with probability 0.05. The middle panel has the same assumptions as the left but $\rho \sim \mathcal{N}(0.5, 0.5)$. The right has the same assumptions as the middle but $\eta \sim \mathcal{U}(0, 1.1)$ and γ is drawn from a density such that $\gamma \in [-0.1, 0]$ implies $g(\gamma) = 9.4$, $\gamma \in (0, 0.45]$ implies $g(\gamma) = \frac{1}{45}$, $\gamma \in (0.45, 0.54]$ implies $g(\gamma) = \frac{5}{9}$, and $g(\gamma) = 0$ for all $\gamma \notin [-0.1, 0.54]$.

ing and enforcement behaviors co-exist for the same parameter values. This means that organizations can differ widely in terms of both the accuracy of reports as well as their enforcement behavior, and indeed expectations *determine* outcomes. Several studies point out the importance of "culture" in law enforcement organizations. By culture, this line of work often refers to the importance of leadership and managerial expectations for the workings of police departments (Cordner 2017; Ingram, Terrill and Paoline 2018; Johnson 2015). Put differently, expectations about proper behavior have a causal effect on what organizational members are doing on the job (see also Schneider and Bose 2017). Such a mechanism is often discussed in the study of police force, i.e., how expectations by supervisors to approve of using coercion to stop perpetrators of crime create incentives of on the ground agents to indeed appropriately apply force (e.g., Ingram, Terrill and Paoline 2018; Terrill, Paoline and Manning 2003). Similarly, in this model, the existence of expectations about the quality of police records can create the possibility of distinct behavioral patterns with respect to both record keeping and enforcement effort.

To see when multiple equilibria may arise, differentiating Λ reveals that

$$-g(1 - F(\hat{\rho})) - f(\hat{\rho})\,[\beta + \Psi] + G(1 - F(\hat{\rho}))\frac{d\Psi}{d\hat{\rho}}.$$

A standard sufficient condition for uniqueness is that Λ is decreasing in the threshold $\hat{\rho}$. The first expression in the preceding expression is indeed negative, representing the standard effect that more effort decreases illicit activity, which decreases the benefits of exerting effort. However, the second expression can be positive or negative. Recall from the previous subsection that the sign of total effect of expected effort on Ψ is determined by the sign of the total, weighted effect on the manipulation stakes:

$$\text{sign}\left(\frac{d\Psi}{d\hat{\rho}}\right) = \text{sign}\left(\frac{\partial\overline{\mu}}{\partial E^s} - \frac{\partial\overline{\mu}}{\partial C^s} g(1 - F(\hat{\rho}))\right).$$

As a result, if Equation 3 admits multiple solutions, it must the be case that when the agent considers increasing effort (marginally increasing $\hat{\rho}$), the reporting stage further incentivizes higher effort, so that $\frac{d\Psi}{d\hat{\rho}}$ is positive. As demonstrated, this can only be the case if (marginally) higher expected effort increases the manipulation stakes, $\bar{\mu}$. Thus, expected effort increasing the incentives to manipulate data is a necessary condition for equilibrium multiplicity.

Proposition 2. *Suppose there are multiple equilibria. Then, for some effort thresholds $\hat{\rho}$, an increase in expected effort increases the stakes of the reporting stage:*

$$\frac{\partial\bar{\mu}}{\partial E^s} - g(1 - F(\hat{\rho}))\frac{\partial\bar{\mu}}{\partial C^s} > 0.$$

In Appendix F, we detail two sufficient conditions for uniqueness. The first follows a standard approach—e.g., Baliga and Sjostrom (2009)—by ensuring that there is enough uncertainty over the private information in the enforcement game. For example, consider Figure 1's middle panel. Here, effort costs are drawn from a Normal distribution with standard deviation 0.5, which has more uncertainty than the left panel where the standard deviation parameter is 0.1. In this case Λ will be increasing in $\hat{\rho}$ with a sufficiently shallow slope. The second ensures that Λ is strictly decreasing in $\hat{\rho}$ over an appropriate interval, thereby ensuring that the right-hand and left-hand sides of Equation 3 are strictly decreasing and increasing in $\hat{\rho}$, respectively. This condition is illustrated in Figure 1's right panel.

4 Empirical Implications

4.1 Data manipulation costs and measurement error

We now examine our key quantities of measurement error and difference in crime coverage. In equilibrium, they are intimately linked. In fact, in this version of the model, they are the same.

Remark 1. *In a full-support equilibrium, $D^s = M^s$.*

As a result, our insights apply to both quantities. Intuitively, Remark 1 follows from two reasons. First, we employ binary variables which are both characterized by a single quantity: the probability of success. Second, because the agent wishes to signal high effort, she never misreports after the no crime outcome, i.e., $R_0^s = 0$. This is important because M^s is increasing in R_0^s (average misclassification is increasing in misclassification after each outcome) while D^s is decreasing in R_0^s (the probability of a reported crime outcome is increasing in R_0^s, which enters D^s negatively). Since $R_0^s = 0$, these differences in the computation of M^s and D^s do not matter, and they are equal to each other.

By Definition 2, crime is under-reported in equilibrium:

$$\begin{aligned} D^s &= \Pr(x = 1|s) - \Pr(\tilde{x} = 1|s) \\ &= \underbrace{(1 - E^s)C^s[E^s]}_{\Pr(x=1|s)} \underbrace{R_1^s[E^s, C^s]}_{\Pr(\tilde{x}\neq x|x=1,s)} > 0. \end{aligned}$$

Moreover, note that D^s and M^s are both a function of enforcement via the probability that the agent finds herself in a situation in which she is tempted to lie, $\Pr(x = 1|s) = (1-E^s)C^s$, and a function of the misreporting probability in this situation, R_1^s. In equilibrium, these two quantities are linked. Nevertheless, to build intuition, it is useful to first examine measurement error when the link between behavior and reporting is acknowledged, i.e., both effort E^s and illicit activity C^s affect R_1^s, but the link between between the agent's and the target's behavior is ignored, i.e., one can vary E^s and C^s independently.

Proposition 3. *Fixing agent effort (E^s), increasing criminal behavior (C^s) increases measurement error. Fixing criminal behavior, increasing agent effort decreases measurement error if and only if the increase in misreporting is sufficiently small:*

$$h(\hat{\eta}^*)\frac{\partial\hat{\eta}^*}{\partial E^s} < \frac{H(\hat{\eta}^*)}{1 - E^s}.$$

Proposition 3 demonstrates the complexity of the forces driving measurement error and the difference in crime coverage. In particular, even if one was able to independently manipulate agent effort, the implications are not clear. Intuitively, agent effort decreases actual crime, which makes it less likely that the agent finds herself in a situation in which she faces incentives to manipulate data. However, when the agent is indeed in this situation, higher effort means more misreporting because the agent has more credibility. Of course, in equilibrium, higher agent effort also decreases illicit activity, which also affects measurement error. Hence, equilibrium effects are even more complex.

Proposition 3 also shows that if one were able to directly decrease the probability of misreporting *without affecting the agent's effort or the target's choice of illicit behavior*, measurement error would decrease. One idea is to increase the agent's manipulation cost. A variety of policies or procedures can be interpreted as increasing manipulation cost, from oversight (Cook and Fortunato 2022) to body cameras (Yokum, Ravishankar and Coppock 2019). To incorporate this in the model, consider a parameter $\sigma \geq 0$, and assume that manipulation costs are now given by $H_\sigma(\eta) \equiv H(\eta - \sigma)$, with support $\left[\underline{\eta} + \sigma, \overline{\eta} + \sigma\right]$.[19] An increase in σ corresponds to a first-order stochastic dominance shift. We now investigate how measurement error M^s responds. To begin with, write measurement error explicitly as a function of σ:

$$M^s(\sigma) = (1 - E^s(\sigma))C^s[E^s(\sigma)]R_1^s\left[E^s(\sigma), C^s[E^s(\sigma)], \sigma\right],$$

where we emphasize that σ affects E^s via Ψ and C^s indirectly via E^s. We have:

$$\frac{\partial M^s}{\partial \sigma} = \underbrace{\frac{\partial E^s}{\partial \sigma} R_1^s \left[-C^s + (1 - E^s)\frac{\partial C^s}{\partial E^s}\right]}_{\text{Effect on Lying State}} + \underbrace{C^s \left[\overbrace{\frac{\partial R_1^s}{\partial \sigma}}^{\text{Direct Effect}} + \frac{\partial E^s}{\partial \sigma}\overbrace{\left(\frac{\partial R_1^s}{\partial E^s} + \frac{\partial R_1^s}{\partial C^s}\frac{\partial C^s}{\partial E^s}\right)}^{\text{Total Effect of Effort}}\right]}_{\text{Effect on Reporting}}. \tag{4}$$

There are several effects. To begin with, there is a direct effect: all else equal in enforcement, it can be shown that as σ increases, the agent manipulates less: $\frac{\partial R_1^s}{\partial \sigma} < 0$. This is the Direct Effect in Equation 4.

Additionally, the change of manipulation costs has spillover effects for the enforcement stage. In particular, the agent's equilibrium enforcement effort will vary with manipulation costs, i.e., $\frac{\partial E^s}{\partial \sigma} \neq 0$. This leads to indirect effects in Equation 4. The next result demonstrates that, under some conditions, an increase in σ increases effort.[20]

Proposition 4. *Assume that H_σ is the Uniform distribution over $\left[\underline{\eta} + \sigma, \overline{\eta} + \sigma\right]$, some types of targets will never choose crime ($G(1) < 1$), and that the agent's effort cost are sufficiently noisy ($F(G(1)\beta) > 0$ and $F(G(0)(\beta + 1)) < 1$). If (s, b) is a full-support equilibrium such that $\hat{\rho}^*$ is the unique solution to Equation 3 and $\Lambda'(\hat{\rho}^*) \neq 1$, then an increase in manipulation costs σ:*

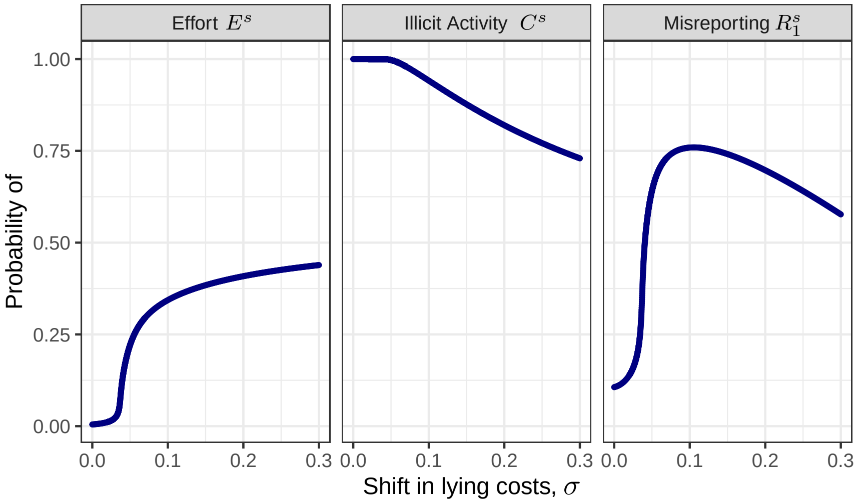
- *decreases the equilibrium probability of data manipulation, R_1^s, when E^s and C^s are held fixed,*
- *increases equilibrium effort E^s,*
- *and decreases equilibrium illicit activity C^s.*

To understand the consequences of Proposition 4 for measurement error, recall that the probability of a crime outcome is $\Pr(x = 1 \mid s) = (1 - E^s)C^s$. Increases in manipulation costs will increase effort and decrease the rate of criminal activity, as stated in Proposition

[19] This specification is from Benabou and Tirole (2011).

[20] The technical condition on the derivative of Λ is necessary for the equilibrium to be well-behaved or "regular" in the game-theoretic sense, e.g., it is a necessary condition to apply the Implicit Function Theorem to compute $\frac{\partial \hat{\rho}^*}{\partial \sigma}$. It is also a generic condition in that every equilibrium threshold $\hat{\rho}^*$ will satisfy the condition for all values of $\beta > 0$ except for at most a closed, Lebesgue-measure-zero subset of $(0, \infty)$.

Figure 2: Equilibrium quantities and data manipulation costs.



Notes: Example with $\beta = 0.2$, $\rho \sim \mathcal{N}(0.5, 0.1)$, $\gamma \sim \mathcal{N}(0.5, 0.1)$, and $\eta \sim \mathcal{U}(\sigma, 0.4 + \sigma)$, where σ represents the severity of lying costs.

4. This is the indirect Effect on the Lying State in Expression 4, where an increase in σ decreases the likelihood of being in the state of the world in which the agent lies, which further decreases measurement error.

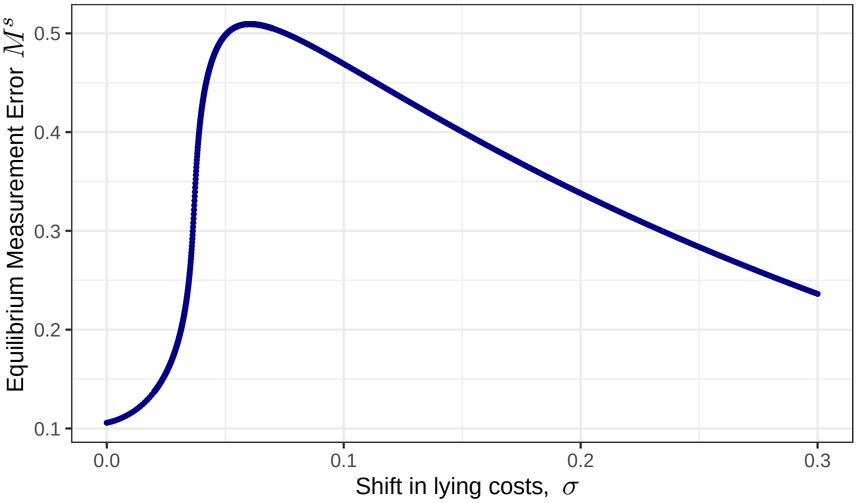
However, there is a final indirect effect in which the enforcement stage again affects behavior at the reporting stage. In particular, more expected effort increases incentives to reclassify—this is the Total Effect of Effort in Equation 4 and it can be positive or negative. If it is positive and large in magnitude, it can overcome the other, negative effects. Hence, an increase in σ can increase or decrease measurement error M^s.

To illustrate these results, consider an example in which the agent's motivation to stop crime, β, is equal to 0.2, effort costs ρ are drawn from $\mathcal{N}(0.5, 0.1)$, opportunity cost γ are drawn from $\mathcal{N}(0.5, 0.1)$, and the data manipulation costs η are drawn from $\mathcal{U}(\sigma, 0.4 + \sigma)$. We vary $\sigma \in [0, 0.3]$ to represent changes in the severity of lying costs.

Figure 2 graphs the equilibrium rates of effort, criminal activity, and misreporting at different levels of lying costs. Notice that there is a unique equilibrium here, so larger lying costs imply more enforcement effort and less criminal activity, illustrating Proposition 4.

Figure 3 shows how these equilibrium quantities map onto measurement error M^s and the difference in crime coverage D^s at various levels of the severity of data manipulation costs, σ. The intuition is that essentially the second indirect effect dominates in this example. Increasing the distribution of lying costs, incentivizes effort in the enforcement stage which implies that criminal activity decreases and measurement error decreases. The agent has a better reputation, which incentivizes the agent to lie in the reporting stage. Hence, equilibrium measurement error increases. When lying costs are small the latter effect dom-

Figure 3: Measurement error and data manipulation costs.



Notes: Example generated using the assumptions in Figure 2, where σ captures the severity of data manipulation costs. Recall that measurement error is equal to the difference in crime coverage, i.e., $M^s = D^s$.

inates; as they get larger, the former (and direct effect) does.

4.2 Measurement error when effort is necessary to detect crime

We briefly discuss insights from a version of the model in which effort positively affects the crime statistic, i.e., $x = ec$. As discussed above, this likely captures crime statistics like speeding tickets, where enforcement effort is required to detect crime. Focusing again on equilibria in which both reports are sent with positive probability, we can show that measurement error, M^s, is closely connected to the difference in crime probabilities, D^s. Here, $D^s = -M^s$. Moreover, we can replicate the characterization of measurement error as the product of the likelihood of the lying state and the misreporting probability:

$$D^s = -M^s = -\underbrace{(1 - E^s C^s[E^s])}_{\Pr(x=0|s)} \underbrace{R^s_0\left[E^s, C^s\right]}_{\Pr(\tilde{x}\neq x|x=0,s)} \leq 0.$$

Thus, in this version of the model, there is over-reporting of crime because the agent wishes to signal that effort was provided, which is necessary to uncover crime.

Proposition 5. *If the actual law enforcement statistic is produced by $x = (1-e)c$, and both reports are sent in equilibrium, then there is under-reporting and measurement error M^s is weakly positive. If the actual law enforcement statistic is produced by $x = ec$, and both reports are sent in equilibrium, then there is over-reporting and measurement error M^s is weakly negative.*

Hence, the nature of enforcement effort determines the sign of the difference between true and reported crime probabilities: preventative enforcement creates incentives to under-report whereas remedial enforcement creates incentives to over-report. Although the over-reporting of crime might seem counterintuitive, one prominent example illustrates the result. According to an external audit, Connecticut state police falsified records of tens of thousands of traffic tickets (CT Insider 2023). The fake tickets were not actually issued, so the reports were only internal and no money was collected. In this case, the over-reporting was only uncovered after an initial investigation found that "four troopers had collectively entered at least 636 fake tickets into the state police computer system over a nine-month stretch to make it appear they were more productive than they actually were" (CT Insider 2023). The rewards were internal involving better assignments, pay increases, promotions and specialty vehicles. These patterns appear in our model where over-reporting should be most likely when examining crime statistics that positively signal enforcement effort, such as ticketing or other remedial policing.

In Appendix E, we characterize full-support equilibria with remedial policing. As in the our characterization above, we show that equilibrium behavior in the enforcement game and in the reporting game are jointly determined, where misreporting affects the incentives for effort and enforcement effort determines the incentives for misreporting. Thus, our insights are robust to a range of law enforcement activities, and hence to a range of empirical studies.

4.3 Causal effects with misreported crime

We now investigate the issue of how the frequency of crime changes as parameters change.

Definition 4. *Given an equilibrium (s, b) parameterized by θ, the treatment effect of θ on the true crime statistic is:*

$$\frac{d\Pr(x=1 \mid s)}{d\theta}.$$

The treatment effect of the parameter θ on the observed crime statistic is:

$$\frac{d\Pr(\tilde{x}=1 \mid s)}{d\theta}.$$

The total derivatives in Definition 4 capture the relationship between how the exogenous parameter θ affects the equilibrium choice probabilities via the characterization in Proposition 1, thereby affecting the true and reported crime rates. We are interested in analyzing the conditions under which these two quantities are the same, or whether analysts under- or overestimate treatment effects.

Definition 5. *Given an equilibrium (s, b), the treatment effect on the observed statistic overestimates the treatment effect on the true statistic if*

$$\frac{d\Pr(\tilde{x}=1 \mid s)}{d\theta} > \frac{d\Pr(x=1 \mid s)}{d\theta}.$$

If the inequality is reversed, then the treatment effect on the observed statistic underestimates the treatment effect on the true statistic.

Note that Definitions 4 and 5 consider small changes in a parameter θ. An alternative approach would consider a discrete change in the parameter θ and compute the difference in the probability of the true and observed crime statistic, respectively. To see this, consider parameter values θ' and θ'' with associated with associated equilibrium profiles s' and s'', respectively. The treatment effects on the true and observed crime statistic are:

$$\Delta_{s'}^{s''} = \Pr\left(x = 1 \mid s'\right) - \Pr\left(x = 1 \mid s''\right) \quad \text{and} \quad \tilde{\Delta}_{s'}^{s''} = \Pr\left(\tilde{x} = 1 \mid s'\right) - \Pr\left(\tilde{x} = 1 \mid s''\right). \tag{5}$$

Employing these definitions, we first show that the bias of a causal estimate, i.e., the difference between these two quantities, is closely connected to measurement error and hence the difference in crime coverage.

Proposition 6. *When measuring the effect of parameter θ on crime, the observed effect is equal to the true effect plus a bias term, where the bias term is the negative of the effect of θ on measurement error, i.e.,*

$$\frac{d\Pr(\tilde{x}=1|s)}{d\theta} = \frac{d\Pr(x=1|s)}{d\theta} + bias, \quad where \quad bias = -\frac{dM^s}{d\theta}.$$

This is intuitive as the only friction between these quantities is imperfect measurement. The result has major implications, however. If a parameter θ can increase or decrease measurement error—such as the lying costs parameter in the previous subsection—it will also increase or decrease the difference in treatment effects. The next Corollary illustrates this a bit differently.

Corollary 1. *The observed treatment effect can be written as an attenuated true treatment effect and a weighted effect on the misreporting probability R_1^s:*

$$\frac{d\Pr(\tilde{x}=1|s)}{d\theta} = (1 - R_1^s)\frac{d\Pr(x=1 \mid s)}{d\theta} + \Pr(x=1 \mid s)\frac{dR_1^s}{d\theta} \tag{6}$$

Corollary 1 shows that the *sign* of $\frac{dR_1^s}{d\theta}$ can be used to sign the bias from strategic misreporting:

- If $\frac{d\Pr(x=1|s)}{d\theta} > 0$ and $\frac{dR_1^s}{d\theta} < 0$, then the observed effect underestimates the true effect.
- If $\frac{d\Pr(x=1|s)}{d\theta} < 0$ and $\frac{dR_1^s}{d\theta} > 0$, then the observed effect overestimates the true effect.

Some empirical studies attempt to directly measure the misreporting probability R_1^s (e.g., Luh 2022). Combined with a conjecture about the sign of the true treatment effect, an examination of how that this variable changes with the treatment can be used to provided a guess on how measurement error changes treatment effects.

We illustrate the content of Proposition 6 with an application to the literature on the effect of economic opportunity costs on crime—for a recent review of the evidence, see Draca and Machin (2015). To investigate this questions, researchers estimate the following canonical regression:[21]

$$\mathbb{E}\left[\tilde{x}_{nt}\right] = \alpha_0 + \alpha_1 \text{EconCond}_{nt} + \text{Controls}_{nt} + \varepsilon_{nt}, \tag{7}$$

where n is an index for the relevant units (districts, states, countries), t are relevant time periods if present (days, months, years), Controls_{nt} are control variables, ε_{nt} is an error term, and EconCond_{nt} is the variable of interest, measuring the economic conditions of the citizens (i.e., their opportunity costs) who potentially engage in illicit activity. The parameter α_1 is the coefficient of interest, measuring the treatment effect of economic conditions.[22]

Our results imply that even if economic conditions were *randomly assigned to units*, the coefficient α_1 does not recover the accurate treatment effect. To see this, note that the left-hand side of Equation 7 is just $\Pr(\tilde{x} = 1 \mid s)$ and consider the following setup. We assume that the agent's motivation, $\beta = 1$, the effort costs ρ are drawn from $\mathcal{N}(0.5, 0.1)$, the data manipulation cost η are drawn from $\mathcal{U}(0, 1.25)$, and the opportunity cost for illicit behavior γ are drawn from $\mathcal{N}(\xi, 1)$. We vary the mean, ξ, to represent exogenous changes to economic opportunities. When ξ is larger, the economy is better, so the target faces higher expected opportunity costs when engaging in illicit activity.

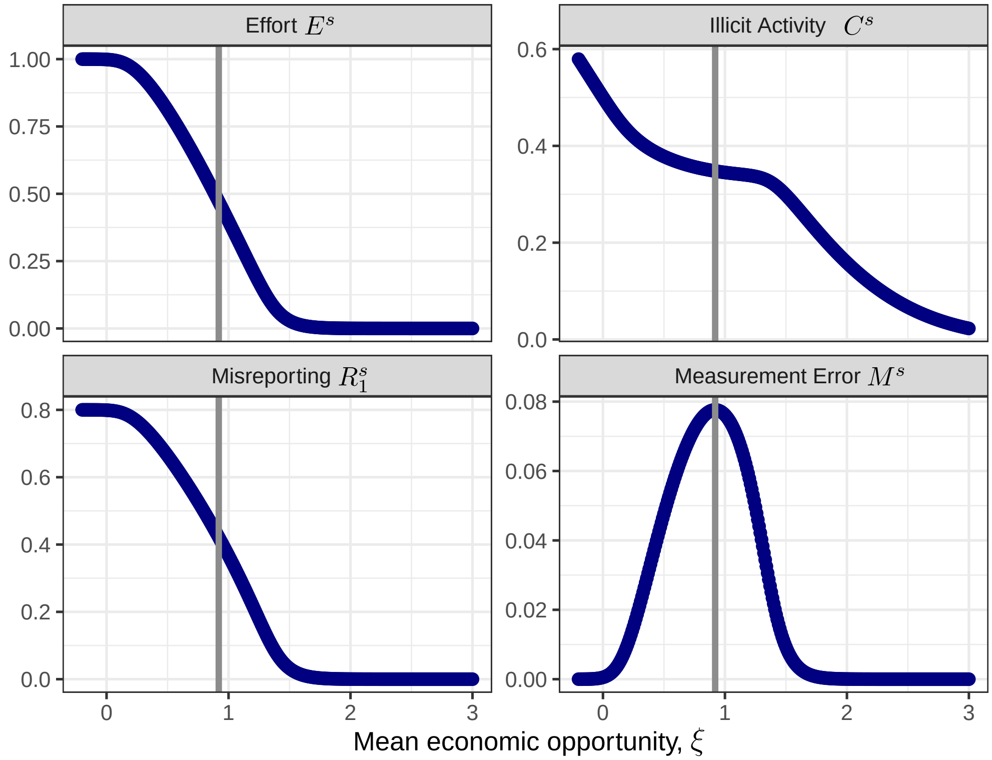
Figure 4 graphs the equilibrium quantities of interest as a function the parameter ξ. Notice that as economic opportunity increases, criminal activity decreases, which is the first-order effect. This leads to reductions in enforcement effort, which subsequently reduces misreporting as sending the no-crime report becomes less believable when effort is smaller. Finally, measurement error is largest when $\xi \approx 0.91$. At this moderate level of economic opportunity, the likelihood of a the crime outcome $\Pr(x = 1|s)$ is not too small, and the misreporting rate, R_1^s, is still substantial.

Figure 5 shows how these equilibrium quantities map onto crime outcomes and treatment effects. The left panel graphs the probability that the enforcement game produces a law enforcement statistic with crime (in blue) and then the probability that the agent reports crime (in orange). Notice the orange line is weakly smaller than the blue line, reflecting

[21] To be precise, most scholarship utilizes data on the number of reported illicit activities in a given location, per time unit. However, this is a consequence of data availability, not a substantive difference: in principle, for sufficiently fine-grained location and time data, at most a single crime can occur. It is also straightforward to generate count data from our model of individual encounters. To do so, denote by $\tilde{X}$ the number of crimes per time and location, and assume it is generated from a Binomial distribution with n possible encounters and success probability $\tilde{p}_s \equiv \Pr(\tilde{x} = 1|s)$. Of course, this assumes that the number of possible encounters is exogenous and that parameter values are the same for all possible encounters in a given unit. However, it is straightforward to introduce additional randomness by varying other parameter values. Similarly, the number of true crimes in the location is given by the variable X, which is given by a binomial distribution with $p_s \equiv \Pr(x = 1|s)$.

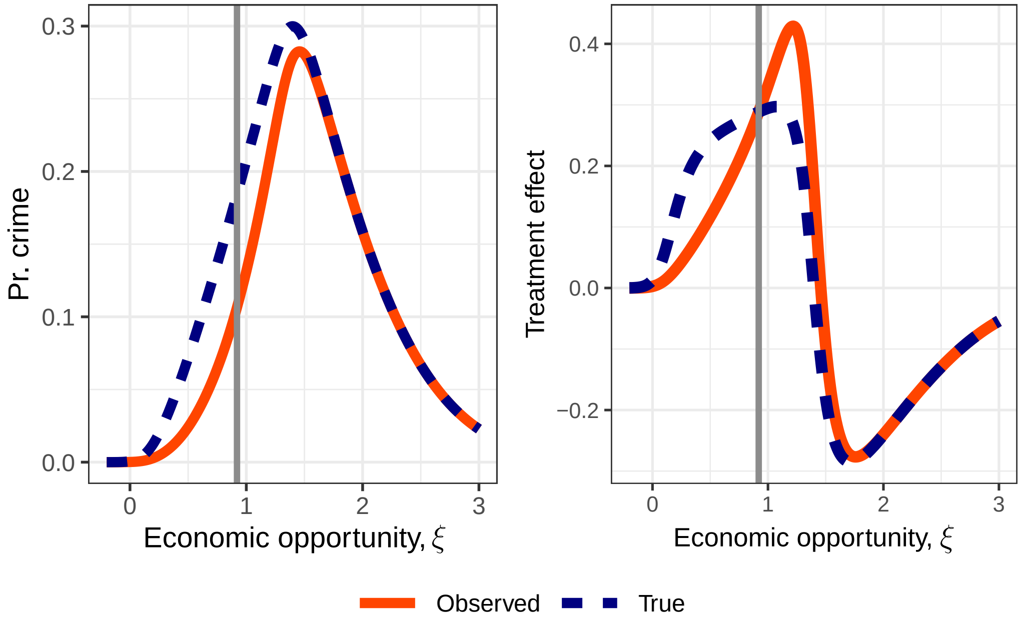
[22] At times, researchers focus on the crime rate, e.g., Bell, Bindler and Machin (2018).

Figure 4: Equilibrium quantities as a function of economic opportunity



Notes: Example generated assuming $\gamma \sim \mathcal{N}(\xi, 1)$, $\beta = 1$, $\rho \sim \mathcal{N}(0.5, 0.1)$, and $\eta \sim \mathcal{U}(0, 1.25)$. The parameter ξ represents economic opportunity, which increases the expected opportunity costs of crime. Grey line highlights the value that maximizes measurement error.

Figure 5: Outcomes and treatment effects as a function of economic opportunity



Notes: Left panels graphs the true crime probability $\Pr(x = 1|s)$ in blue and the observed crime probability $\Pr(\tilde{x} = 1|s)$ in orange as a function of economic opportunity, ξ. Right panel graphs the true treatment effect $\frac{d\Pr(x=1|s)}{d\xi}$ in blue and the observed treatment $\frac{d\Pr(\tilde{x}=1|s)}{d\xi}$ in orange. Example generated using same assumptions as in Figure 4.

the fact that crime is under-reported in equilibrium, i.e., $M^s > 0$. Even though the rate of illicit activity decreases in ξ, this panel shows that the probability of the crime outcome is a non-monotonic function of ξ. This non-monotonicity arises because enforcement effort also decreases with greater economic opportunity, ξ.

The right panel in Figure 5 then graphs the true treatment effect $\frac{\partial \Pr(x=1)}{\partial \xi}$ in blue and the observed treatment effect $\frac{\partial \Pr(\tilde{x}=1)}{\partial \xi}$ in orange. The difference between these two quantities can be interpreted as bias from strategic misreporting. Notice that the observed treatment effect can either underestimate ($\xi < 0.91$) or overestimate ($\xi > 0.91$) the true treatment effect. As Proposition 6 demonstrates, the effect is underestimated when increasing ξ increases measurement error. The effect is overestimated when increasing ξ reduces measurement error.

Another pattern is that the difference in treatment effects is not a monotonic function of measurement error. From the statistical literature on measurement error in the *in*dependent variable, one can gain the intuition that as measurement error increases, (attenuation) bias becomes more severe. In Figure 5, there is no difference between treatment effects (bias) precisely when measurement error is maximal ($\xi \approx 0.91$). This illustrates a consequence of Proposition 6. When measurement error is at a local maximum, then $\frac{\partial M^s}{\partial \theta} = 0$, there is no bias as the difference in treatment effects is zero.

Finally, notice that maximal measurement error is not a necessary condition for no bias. When economic opportunity is very small ($\xi < 0$) or large ($\xi > 1.75$), the difference between the treatment effects is essentially zero. To see why, recall the discrete definition of treatment effects in Equation 5. Plugging in the relevant quantities and simplifying shows that

$$\tilde{\Delta}_{s'}^{s''} - \Delta_{s'}^{s''} = M^{s'} - M^{s''}, \tag{8}$$

i.e., the difference in treatment effects is equal to the difference in measurement errors. Thus, when neither profile (s' or s'') features measurement error, the treatment effects are equal, which is precisely the situation when ξ takes extreme values in Figures 4 and 5. Conversely, $\tilde{\Delta}_{s'}^{s''} > \Delta_{s'}^{s''}$ if and only $M^{s'} > M^{s''}$. When $M^{s'} > M^{s''}$, the observed effect overestimates the true effect because it captures a decrease in under-reporting, and hence more reported crime.

5 Discussion and Conclusion

5.1 Relationship to Existing Literature

In this section, we discuss our contribution in relation to the existing literature on measurement error and its consequences for inference. Scholars are well aware that many variables in the social sciences suffer from measurement error, especially variables collected via surveys, and have made progress in understanding the consequences of measurement error for empirical analysis (for an overview, see Bound, Brown and Mathiowetz 2001). One important insight is that with a binary dependent variable, measurement error is always non-classical and can seriously affect inferences about parameters (Bound, Brown and Mathiowetz 2001; Meyer and Mittag 2017). Our results contribute to this line of work by demonstrating that a particular kind of measurement error is in fact generic for enforcement data.

To see this, consider this canonical setup. Let n index observations $1, \dots, N$, where N is the sample size. The true data generating process, as a function of a parameter θ_n, is assumed to be $x_n = \theta_n'\beta + \varepsilon_n$, where ε_n is the error term. F is the CDF of $-\varepsilon_n$. As in the example in Section 4.3, we interpret θ_n to be a policy relevant variable such as economic opportunity. The true outcome is given by the variable $x_n \in \{0, 1\}$ such that $\Pr(x_n = 1 \mid \theta_n) = F(\theta_n'\beta)$. However, the analyst is assumed to only observe the variable $\tilde{x}_n \in \{0, 1\}$. A key ingredient in this setup is interim misclassification probabilities, referred as misreporting in our model. Using our notation—but omitting the reference to a strategy profile s since these are not endogenously derived from equilibrium behavior—the interim misclassification probabilities are

$$\Pr(\tilde{x}_n = 1|x_n = 0) = R_0(\theta_n) \quad \text{and} \quad \Pr(\tilde{x}_n = 0|x_n = 1) = R_1(\theta_n). \tag{9}$$

An important distinction is whether these depend on the value of the covariate of interest, θ_n. If this is not the case, so that $\frac{\partial R_0}{\partial \theta_n} = 0$ as well as $\frac{\partial R_1}{\partial \theta_n} = 0$, then misclassification is considered to be "conditionally random" (Hausman, Abrevaya and Scott-Morton 1998). By contrast, if the misclassification probabilities do depend on the covariate θ_n, then misclassification is not conditionally random and the partial derivatives of R_0 and R_1 with respect to θ_n are not 0. These situations are depicted graphically in Figure 6.

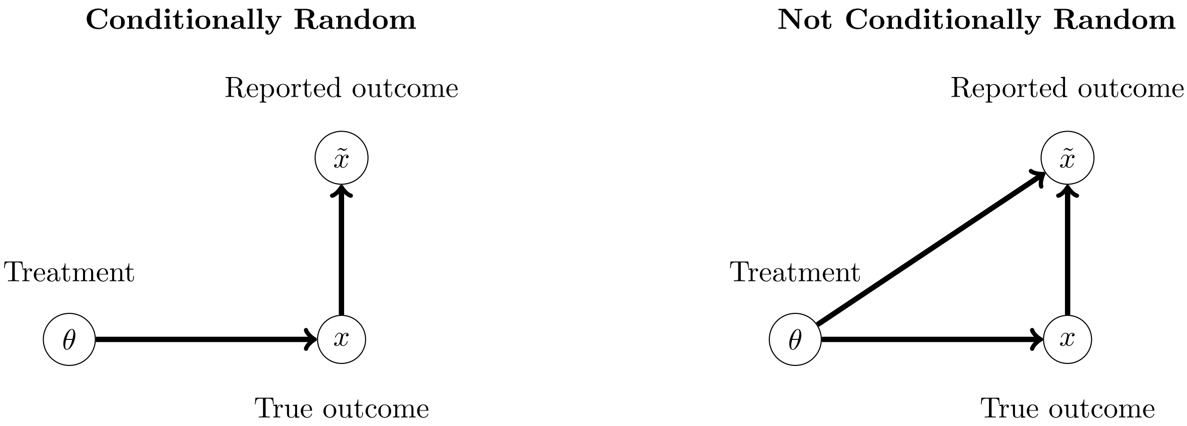


Figure 6: Existing approaches to measurement error.

Both types of misclassification can lead to bias. To see this, recall that a standard result for non-linear models is that the true marginal effect for covariate θ_n is

$$\frac{\partial \Pr(x_n = 1 \mid \theta_n)}{\partial \theta_n} = f(\theta_n' \beta)\beta.$$

However, the observed marginal effect is

$$\frac{\partial \Pr(\tilde{x}_n = 1 \mid \theta_n)}{\partial \theta_n} = \underbrace{\frac{\partial R_0(\theta_n)}{\partial \theta_n}}_{\equiv E_1} - \underbrace{\left(\frac{\partial R_0(\theta_n)}{\partial \theta_n} + \frac{\partial R_1(\theta_n)}{\partial \theta_n}\right) F(\theta_n'\beta)}_{\equiv E_2} + \underbrace{(1 - R_0(\theta_n) - R_1(\theta_n)) f(\theta_n'\beta)\beta}_{\equiv E_3}. \tag{10}$$

When misclassification is conditionally random, then $\frac{\partial R_k(\theta_n)}{\partial \theta_n} = 0$, for $k = 1, 2$, which implies (i) $E_1 = E_2 = 0$ and (ii) the observed marginal effect is attenuated (closer to zero) if $R_0 + R_1 < 1$. The situation is even more complex and challenging if measurement error directly depends on the value of the treatment. By inspection, the terms E_1 and E_2 can either be negative or positive. As a consequence, when the rate of classification depends on the observable variable, we cannot sign the bias from comparing the true marginal effect and the observed marginal effect even if $R_0(\theta_n) + R_1(\theta_n) < 1$.

An example of this kind of pernicious measurement error is discussed in Weidmann (2016) who studies the effects of cellphone coverage on violence. In that paper, cellphone coverage has two effects. The first effect is that cellphones may help groups overcome collective action problems to mobilize for violence (Pierskalla and Hollenbach 2013). The second is that cellphones directly effect the reporting of violence (Dafoe and Lyall 2015). Weidmann (2016) shows that the second effect means that cellphone coverage directly influences the misreporting rate of violence, making it difficult to estimate the first effect of cellphone coverage on violence. Similarly, in Meyer and Mittag (2017), the misclassification probabilities can be different for different units and dependent on the value of a covariate. Finally, in Blattman et al. (2016), an observed variable is assumed to be a linear function of the true variable and the treatment variable, implying that the treatment affects misreporting, conditional on the true variable.

These data generating process are plausible and allow researchers to derive formulas for biases in parameter estimates, often within a regression framework. For example, Bound, Brown and Mathiowetz (2001) and Meyer and Mittag (2017) derive regression analogues of our Proposition 6, showing that within their linear setup, the bias is equal to a regression of measurement error on the covariate. Besides characterization of bias, a regression framework also allows the development of a "diagnostic procedure" (Weidmann 2016) to test if biased misreporting is an issue.[23] However, in these cases, the data generating process connecting the treatment, the true dependent variable, and the observed dependent variable is entirely assumed and judgment on whether a variable directly affects the misreporting probability is based on how plausible it is that the covariate in question, such as cellphone coverage, directly affects misreporting.

By contrast, we assume a particular strategic situation between a citizen and an agent, but then *derive* the data generating process connecting the treatment variable to the true and observed outcome variables. Our model shows that the right panel in Figure 6 is in fact the generic kind for enforcement agencies, as show in Figure 7.

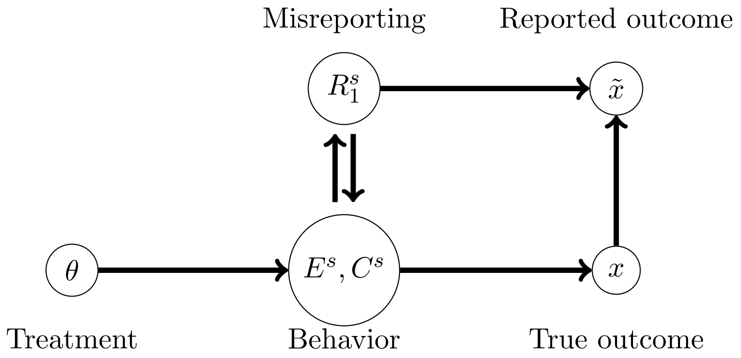


Figure 7: Measurement error with endogenous behavior and misreporting.

As the Figure shows, in our model, a treatment is connected to both the true outcome

[23] Weidmann (2016) notes that datasets often contain information about violence severity, and argues that high-severity events are less likely to be misreported. Hence, one should rerun the regression of cellphone coverage on violence for different subsets of severity and assess whether results differ.

and the observed outcome through our endogenous choices. Hence, for enforcement agencies who are in charge of maintaining records, *if* a treatment has a causal effect on the true outcome, it *must* have an effect on the misreporting probability. The reason is that reporting is optimally conditioned on (expectations of) enforcement behavior. Hence, if a variable affects behavior (and hence the true variable), it must also affect the misreporting probability. Formally, in our model, the true outcome variable's distribution is $(1 - E^s)C^s$. The misreporting probability is $R_1^s[E^s, C^s]$, with $\frac{\partial R_1^s}{\partial E^s} > 0$ and $\frac{\partial R_1^s}{\partial C^s} > 0$. Thus, misreporting depending on treatment status occurs whenever the treatment has an effect on the true variable, x. The example in Section 4.3 illustrates this specific problem: economic opportunity has no direct effect on the agent's reporting incentives, but it does directly affect incentives for crime and enforcement. Because economic opportunity affects behavior in the enforcement game, it affects the misreporting (interim misclassification) via the agent's signaling or reputational incentives.

Figure 7 illustrates that the bias is *not* created because of selection or more generally confounding (which are the usual challenges to credible causal inference). In other words, there are no backdoor paths from the treatment θ to either x or $\tilde{x}$. However, the researcher is assumed not to have access to the true outcome variable x but has to rely on the observed variable $\tilde{x}$, and there is an additional path, via the misreporting probability R_1^s, that connects the treatment θ to $\tilde{x}$. Critically, that path runs through the key mediator behavior (E^s and C^s), meaning that the reason misreporting is a problem is the very reason the treatment has an effect on the true outcome.

Returning to Expression 10, but assuming that $R_0(\theta_n) = 0$, which is an equilibrium phenomenon in our model, we have:

$$\begin{aligned}\frac{\partial \Pr(\tilde{x}_n = 1 \mid \theta_n)}{\partial \theta_n} &= -\frac{\partial R_1(\theta_n)}{\partial \theta_n} F(\theta_n' \beta) + (1 - R_1(\theta_n)) f(\theta_n' \beta)\beta \\ &= -\frac{\partial R_1(\theta_n)}{\partial \theta_n} \Pr(x_n = 1 \mid \theta_n) + (1 - R_1(\theta_n)) \frac{\partial \Pr(x_n = 1 \mid \theta_n)}{\partial \theta_n},\end{aligned}$$

where the second equality plugs in the general terms again. Comparing this with Corollary 1, shows that our model creates an inference problem in which the rate of misclassification depends on the treatment, i.e., misclassification is not conditionally random.

This discussion highlights the importance of having an explicit model of enforcement and reporting. Without such a model of misreporting, researchers might be stuck arguing about particular exogenous data generating processes, and whether or not it is plausible that a treatment directly affects the misreporting probabilities R_0 and R_1. This issue affects work that starts from particular "structural" equations or from causal graphs. By contrast, we instead start by assuming a particular kind of strategic interaction between law enforcement agents and citizens, and endogenizing misreporting by enforcement agencies. The key assumption is that agencies internalize a signaling or reputation incentive. Such an incentive likely exists when there is a third party—either internal to the agency like a manager or external like a funding organization—that uses the reports to monitor the agency's behavior. Importantly, this yields an endogenous account of measurement error that, in contrast to the data generating processes in Figure 6 cannot be formulated as a directed acyclic graph (DAG). The reason is that in our model, there exists a reciprocal, equilibrium relationship between behavior and misreporting—as highlighted by the two arrows in Figure 7. As a consequence, while accounts that use DAGs as primitives capture important intuitions and allow researchers to provide a classification of different kinds of measurement error (as in Hernán and Cole 2009; Knox, Lucas and Cho 2022), they cannot capture our data generating process and the kinds of inferential challenges that we highlight.

Our analysis shows why misclassification that is *not* conditionally random is in fact a natural case to consider for data collected by enforcement agencies with signaling or reputational concerns. Moreover, we emphasize that intervention aimed at decreasing measurement error may actually backfire and worsen data quality. This also affects inferences, so that empirical scholarship estimates of treatment effects can be under- or overestimated. Given that scholarship extensively utilizes data by collected by law enforcement agencies to answer research questions, our model provides a framework for thinking about the consequences of such strategic misreporting.

5.2 Implications for Empirical Research

What are the implications of our theoretical model for empirical research? First, concerning descriptive statistics, our analysis implies that it is useful to split up the data by crime categories. In our model, crimes in which agent effort is necessary to uncover wrongdoings (like speeding, or the possession of illegal substances) are over-reported whereas crimes in which agent effort is important to reduce the opportunities for engaging in them (like violent robbery) are under-reported. Importantly, this holds for a given distribution of data manipulation costs, H, whereas, as we have shown, changing this distribution can have counter-intuitive effects on misreporting and data quality. As a result, while it is difficult to order data sets in terms of their quality as a function of (perceived) agency misreporting costs, it is possible to make relative statements about misreporting for different crime categories for a given agency.

Moreover, generally, observing reported crime over time (for a single agency) is not sufficient for learning the rate of misclassification *even if the costs of manipulation are increasing* (cf. Cook and Fortunato 2022). The reason is that, as we have shown, data quality is a complicated function of manipulation costs, and it can decrease if manipulation costs increase.

Second, concerning causal inferences, our main result is a negative one: *if* the agency has any reputation incentives, i.e., cares about its perceived effort, the relationship between the true and the observed treatment effect will be too complex to meaningfully bound treatment effects. In particular, the observed treatment effect can be too large or too small, relative to the true treatment effect, and for a given treatment θ, signing the bias term $-\frac{\partial M^s}{\partial \theta}$ is very difficult. The question is how to best deal with this issue. We discuss two approaches that researchers can pursue.

One way is to gather additional data. For example, our Corollary 1 implies that conjectures about true treatment effects and the total effect of the treatment on misreporting, $\frac{dR_1^s}{d\theta}$, one can, under some conditions, say if the observed treatment effect under- or overestimates the true treatment effect. Making an accurate conjecture about the total effect of the treatment on the misreporting is obviously challenging as well. One solution comes from audit studies in which (more) accurate data is obtained for a subset of units (Blattman et al. 2016; Garbiras-Díaz and Slough 2022). This allows researchers to learn the probability of misreporting and hence one can estimate the effect of the treatment on this probability. The downside, of course, is that such a study is very costly and, partially as a result, sometimes infeasible.

Another way is to deal with misreporting at the design stage, e.g., when designing a field experiment. A broad conclusion from both informal reasoning and strategic information models is that competition can often foster accurate communication (Gentzkow and Shapiro 2008). This suggests that researchers should focus on implementing designs in areas in which *both the treatment and control groups* are characterized by a competitive information environment, either in terms of media, non-governmental organizations, or multiple enforcement agencies (e.g., a Sheriff and the police department). We caution, however, that one would ideally include these organizations as separate players in our model to evaluate whether it is indeed the case that their inclusion decreases misreporting (by any one agency). This is an important avenue for future work on the quality of administrative policing data.

Finally, we emphasize that our model can also be applied to other areas of interests. Consider the following relabeling of our model: the agent is a repressive agent aligned with an authoritarian regime. They can either repress a citizen when the citizen is trying to vote for the opposition candidate or not ("effort"). In addition, they can falsify a vote if it occurs ("misreporting"). The target is a citizen who can either attempt to vote for the opposition candidate or not ("illicit activity"). That candidate's true vote count is x and the observed vote count is $\tilde{x}$. Our results imply that, consistent with conventional wisdom, it is very difficult to learn accurate treatment effects from electoral statistics published by authoritarian regimes.

References

Alonso, Ricardo and Odilon Câmara. 2023. "Organizing Data Analytics." *Management Science* forthcoming.

Antonovics, Kate and Brian G. Knight. 2009. "A new look at racial profiling: Evidence from the Boston Police Department." *Review of Economics and Statistics* 91(1):163–177.

Anwar, Shamena and Hanming Fang. 2006. "An alternative test of racial prejudice in motor vehicle searches: Theory and evidence." *American Economic Review* 96(1):127–151.

Arnold, R. Douglas and Nicholas Carnes. 2012. "Holding mayors accountable: New York's executives from Koch to Bloomberg." *American Journal of Political Science* 56(4):949–963.

Arora, Ashna. 2023. "Juvenile Crime and Under-Recording." Unpublished manuscript. Retrieved December 2023, from `https://www.dropbox.com/scl/fi/zkz7iymzaqnhu43r59ia7/Juvenile_Crime_and_Under_Recording.pd`

Avenhaus, Rudolf, Bernhard Von Stengel and Shmuel Zamir. 2002. Inspection games. Vol. 3 of *Handbook of Game Theory with Economic Applications* North Holland pp. 1947–1987.

Ba, Bocar, Patrick Bayer, Nayoung Rim, Roman Rivera and Modibo Sidib. 2021. "Police Officer Assignment and Neighborhood Crime." Unplished manuscript. Retrieved December 2023, from `https://www.dropbox.com/s/vz4pd7wzhc4ennb/PoliceOfficerAssignment.pdf`.

Baliga, Sandeep and Tomas Sjostrom. 2009. "Conflict Games with Payoff Uncertainty." Unpublished manuscript. Retrieved January 2023, from `https://www.kellogg.northwestern.edu/faculty/baliga/htm/stagchick.pdf`.

Bell, Brian, Anna Bindler and Stephen Machin. 2018. "Crime scars: recessions and the making of career criminals." *Review of Economics and Statistics* 100(3):392–404.

Bell, Brian, Laura Jaitman and Stephen Machin. 2014. "Crime deterrence: Evidence from the London 2011 riots." *Economic Journal* 124(576):480–506.

Benabou, Roland and Jean Tirole. 2011. "Laws and norms." *NBER Working Paper* (17579). **URL:** *https://www.nber.org/papers/w17579*

Berman, Eli, Jacob N. Shapiro and Joseph H. Felter. 2011. "Can hearts and minds be bought? The economics of counterinsurgency in Iraq." *Journal of Political Economy* 119(4):766–819.

Blair, Graeme, Jeremy M Weinstein, Fotini Christia, Eric Arias, Emile Badran, Robert A Blair, Ali Cheema, Ahsan Farooqui, Thiemo Fetzer, Guy Grossman et al. 2021. "Community policing does not build citizen trust in police or reduce crime in the Global South." *Science* 374(6571):eabd3446.

Blair, Robert A., Sabrina M. Karim and Benjamin S. Morse. 2019. "Establishing the rule of law in weak and war-torn states: Evidence from a field experiment with the Liberian National Police." *American Political Science Review* 113(3):641–657.

Blattman, Christopher, Julian Jamison, Tricia Koroknay-Palicz, Katherine Rodrigues and Margaret Sheridan. 2016. "Measuring the measurement error: A method to qualitatively validate survey data." *Journal of Development Economics* 120:99–112.

Bound, John, Charles Brown and Nancy Mathiowetz. 2001. Measurement Error in Survey Data. In *Handbook of Econometrics*, ed. James J. Heckman and Edward Leamer. Vol. 5 Amsterdam: Elsevier.

British Broadcasting Corporation. 2013. "Crime statistics are manipulated, says police chief." `https://www.bbc.com/news/uk-25022680` (accessed April 4, 2023).

Bueno De Mesquita, Ethan and Scott A. Tyson. 2020. "The commensurability problem: Conceptual difficulties in estimating the effect of behavior on behavior." *American Political Science Review* 114(2):375–391.

Charnysh, Volha. 2019. "Diversity, institutions, and economic outcomes: Post-WWII displacement in Poland." *American Political Science Review* 113(2):423–441.

Clark, Tom S, Elisha Cohen, Adam Glynn, Michael Leo Owens, Anna Gunderson and Kaylyn Jackson Schiff. 2020. "Are police racially biased in the decision to shoot?" *Working Paper* .

Cook, Scott J and David Fortunato. 2022. "The Politics of Police Data: State Legislative Capacity and the Transparency of State and Substate Agencies." *American Political Science Review* .

Cordner, Gary. 2017. "Police culture: Individual and organizational differences in police officer perspectives." *Policing: An International Journal of Police Strategies & Management* 40(1):11–25.

CT Insider. 2023. "'High likelihood' hundreds of CT state police troopers falsified thousands of traffic tickets, auditor says." `https://www.ctinsider.com/news/article/ct-state-police-troopers-false-tickets-18162917.php` (accessed July 10, 2023).

Czaika, Mathias and Mogens Hobolth. 2016. "Do restrictive asylum and visa policies increase irregular migration into Europe?" *European Union Politics* 17(3):345–365.

Dafoe, Allan and Jason Lyall. 2015. "From cell phones to conflict? Reflections on the emerging ICT–political conflict research agenda." *Journal of Peace Research* 52(3):401–413.

Dallas Morning News. 2020. "Former Dallas officer who wrote fake traffic tickets sentenced to 3 years probation." `https://www.dallasnews.com/news/courts/2020/09/09/former-dallas-officer-who-wrote-fake-traffic-tickets-sentenced-to-3-years-probation/` (accessed April 4, 2023).

Di Lonardo, Livio and Tiberiu Dragu. 2021. "Counterterrorism Policy in an Uncertain World." *Journal of Politics* 83(4):1857–60.

Di Salvatore, Jessica. 2019. "Peacekeepers against criminal violence—unintended effects of peacekeeping operations?" *American Journal of Political Science* 63(4):840–858.

Draca, Mirko and Stephen Machin. 2015. "Crime and economic incentives." *Annual Review of Economics* 7(1):389–408.

Dragu, Tiberiu. 2011. "Is there a trade-off between security and liberty? Executive bias, privacy protections, and terrorism prevention." *American Political Science Review* 105(1):64–78.

Dragu, Tiberiu and Adam Przeworski. 2019. "Preventive repression: Two types of moral hazard." *American Political Science Review* 113(1):77–87.

Dube, Arindrajit, Oeindrila Dube and Omar García-Ponce. 2013. "Cross-border spillover: US gun laws and violence in Mexico." *American Political Science Review* 107(3):397–417.

Duggan, Mark. 2001. "More guns, more crime." *Journal of Political Economy* 109(5):1086–1114.

Dynes, Adam M. and John B. Holbein. 2020. "Noisy retrospection: The effect of party control on policy outcomes." *American Political Science Review* 114(1):237–257.

Eckhouse, Laurel. 2022. "Metrics Management and Bureaucratic Accountability: Evidence from Policing." *American Journal of Political Science* 66(2):385–401.

Eterno, John A. and Eli B. Silverman. 2017. *The crime numbers game: Management by manipulation*. CRC Press.

Fearon, James D. 1999. "Electoral accountability and the control of politicians: selecting good types versus sanctioning poor performance." *Democracy, Accountability, and Representation* 55:61.

Fox, Justin and Richard Van Weelden. 2012. "Costly transparency." *Journal of Public Economics* 96(1-2):142–150.

Fryer Jr, Roland G. 2019. "An empirical analysis of racial differences in police use of force." *Journal of Political Economy* 127(3):1210–1261.

Garbiras-Díaz, Natalia and Tara Slough. 2022. "The Limits of Decentralized Administrative Data Collection: Experimental Evidence from Colombia." Unplished manuscript. Retrieved June 2023, from `http://www.taraslough.com/assets/pdf/decentralized_data.pdf`.

Gentzkow, Matthew and Jesse M. Shapiro. 2008. "Competition and Truth in the Market for News." *Journal of Economic Perspectives* 22(2):133–154.

Gibilisco, Michael and Jessica Steinberg. 2022. "Strategic Reporting: A Formal Model of Bias in Conflict Events." *American Political Science Review* forthcoming.

Hausman, Jerry A., Jason Abrevaya and Fiona M. Scott-Morton. 1998. "Misclassification of the dependent variable in a discrete-response setting." *Journal of econometrics* 87(2):239–269.

Hernán, Miguel A. and Stephen R. Cole. 2009. "Invited commentary: causal diagrams and measurement bias." *American journal of epidemiology* 170(8):959–962.

Horz, Carlo M and Moritz Marbach. 2022. "Economic Opportunities, Emigration and Exit Prisoners." *British Journal of Political Science* 52(1):21–40.

Hübert, Ryan and Andrew T. Little. 2023. "A Behavioural Theory of Discrimination in Policing." *Economic Journal* forthcoming.

Ingram, Jason R., William Terrill and Eugene A. Paoline. 2018. "Police culture and officer behavior: Application of a multilevel framework." *Criminology* 56(4):780–811.

Jassal, Nirvikar. 2020. "Gender, law enforcement, and access to justice: Evidence from all-women police stations in India." *American Political Science Review* 114(4):1035–1054.

Johnson, Richard R. 2015. "Police organizational commitment: The influence of supervisor feedback and support." *Crime & Delinquency* 61(9):1155–1180.

Kartik, Navin and Richard Van Weelden. 2019. "Informative cheap talk in elections." *The Review of Economic Studies* 86(2):755–784.

Khanna, Gaurav, Carlos Medina, Anant Nyshadham, Christian Posso and Jorge Tamayo. 2021. "Job Loss, Credit, and Crime in Colombia." *American Economic Review: Insights* 3(1):97–114.

Knowles, John, Nicola Persico and Petra Todd. 2001. "Racial bias in motor vehicle searches: Theory and evidence." *Journal of Political Economy* 109(1):203–229.

Knox, Dean, Christopher Lucas and Wendy K Tam Cho. 2022. "Testing causal theories with learned proxies." *Annual Review of Political Science* 25:419–441.

Knox, Dean, Will Lowe and Jonathan Mummolo. 2020. "Administrative records mask racially biased policing." *American Political Science Review* 114(3):619–637.

Kovandzic, Tomislav V., Lynne M. Vieraitis and Denise Paquette Boots. 2009. "Does the death penalty save lives? New evidence from state panel data, 1977 to 2006." *Criminology & Public Policy* 8(4):803–843.

Levitt, Steven D. 2002. "Using electoral cycles in police hiring to estimate the effects of police on crime: Reply." *American Economic Review* 92(4):1244–1250.

Los Angeles Times. 2015. "LAPD underreported serious assaults, skewing crime stats for 8 years." `https://www.latimes.com/local/cityhall/la-me-crime-stats-20151015-story.html` (accessed April 4, 2023).

Luh, Elizabeth. 2022. "Not so black and white: Uncovering racial bias from systematically misreported trooper reports." *Available at SSRN 3357063* .

Magaloni, Beatriz, Edgar Franco-Vivanco and Vanessa Melo. 2020. "Killing in the slums: Social order, criminal governance, and police violence in Rio de Janeiro." *American Political Science Review* 114(2):552–572.

McCall, Andrew. 2019. "Resident assistance, police chief learning, and the persistence of aggressive policing tactics in Black neighborhoods." *Journal of Politics* 81(3):1133–1142.

Meyer, Bruce D and Nikolas Mittag. 2017. "Misclassification in binary choice models." *Journal of Econometrics* 200(2):295–311.

New York Times. 2004. “Union Leaders Allege Fudging Of Statistics On City Crime.” https://www.nytimes.com/2004/03/24/nyregion/union-leaders-allege-fudging-of-statistics-on-city-crime.html?searchResultPosition=9 (accessed January 3, 2023).

Olson, Mancur. 1993. “Dictatorship, democracy, and development.” *American political science review* 87(3):567–576.

Patty, John W and Elizabeth Maggie Penn. 2015. “Analyzing big data: social choice and measurement.” *PS: Political Science & Politics* 48(1):95–101.

Pierskalla, Jan H. and Florian M. Hollenbach. 2013. “Technology and collective action: The effect of cell phone coverage on political violence in Africa.” *American Political Science Review* 107(2):207–224.

Reuters. 2012. “NYPD report confirms manipulation of crime stats.” https://www.reuters.com/article/us-crime-newyork-statistics/nypd-report-confirms-manipulation-of-crime-stats-idUSBRE82818620120309 (accessed April 2, 2023).

Roger, Guillaume. 2013. “Optimal contract under moral hazard with soft information.” *American Economic Journal: Microeconomics* 5(4):55–80.

Schneider, Patrick and Gautam Bose. 2017. “Organizational cultures of corruption.” *Journal of Public Economic Theory* 19(1):59–80.

Shaver, Andrew et al. 2022. “News Media Reporting Patterns and our Biased Understanding of Global Unrest.” (ESOC Working Paper No. 32). Empirical Studies of Conflict Project. Retrieved March 2022, from http://esoc.princeton.edu/wp32.

Slough, Tara. 2023. “Phantom Counterfactuals.” *American Journal of Political Science* 67(1):137–153.

Slough, Tara and Scott A Tyson. 2022. “External Validity and Meta-Analysis.” *American Journal of Political Science* .

Stashko, Allison. 2022. “Do Police Maximize Arrests or Minimize Crime? Evidence from Racial Profiling in U.S. Cities.” *Journal of the European Economic Association* 21(1):167–214.

Terrill, William, Eugene A Paoline and Peter K. Manning. 2003. “Police culture and coercion.” *Criminology* 41(4):1003–1034.

Weidmann, Nils B. 2016. “A closer look at reporting bias in conflict event data.” *American Journal of Political Science* 60(1):206–218.

Yokum, David, Anita Ravishankar and Alexander Coppock. 2019. “A randomized control trial evaluating the effects of police body-worn cameras.” *Proceedings of the National Academy of Sciences* 116(21):10329–10332.

APPENDIX (online only)

A Omitted proofs

A.1 Proof of Lemma 1

Proof. To prove (1), suppose the contrary. That is, suppose there exists equilibrium (s, b) such that $\Pr(\tilde{x} = 1) \in (0, 1)$ and either (i) $R_0^s > 0$ or (ii) $R_1^s = 1$.

First, we maintain (i) and show that a contradiction arises. To do this, consider the reporting subgame after outcome $x = 0$. Because $R_0^s > 0$, some agents must reclassify so $b_1 - \eta > b_0$ for some $\eta \in \text{supp}(H)$. Because $\min \text{supp}(H) = \underline{\eta} \geq 0$, we must have $b_1 > b_0$. Furthermore, $b_1 > b_0$ and $\eta \geq 0$ imply that agents with outcome $x = 1$ will never reclassify, so $R_1^s = 0$.

We must either have $E^s = 0$ or $E^s > 0$. Suppose $E^s = 0$. Because $\Pr(\tilde{x} = 1) \in (0, 1)$, both reports $\tilde{x} \in \{0, 1\}$ are sent with positive probability. As such Bayes rules implies $b_1 = b_0 = 0$, a contradiction.

Suppose $E^s > 0$, then crime outcome $x = 0$ occurs with positive probability. Because

$\Pr(\tilde{x}=1) \in (0,1)$ by assumption, after seeing $\tilde{x}=0$, equilibrium beliefs satisfy Bayes rule:

$$\begin{aligned} b_0 &= \Pr(e=1 \mid \tilde{x}=0) \\ &= \frac{\Pr(\tilde{x}=0 \mid e=1)\Pr(e=1)}{\Pr(\tilde{x}=0)} \\ &= \frac{(1-R_0^s)E^s}{\Pr(\tilde{x}=0 \mid x=0)\Pr(x=0)+\Pr(\tilde{x}=0 \mid x=1)\Pr(x=1)} \\ &= \frac{(1-R_0^s)E^s}{(1-R_0^s)(1-\Pr(x=1))}, \end{aligned}$$

where the last equality follows because $R_1^s = \Pr(\tilde{x}=0 \mid x=1) = 0$. An implication of the above logic is that $\Pr(\tilde{x}=1) \in (0,1)$ implies $R_0^s < 1$. If not, then $\Pr(\tilde{x}=1)=1$, a contradiction. Likewise, after seeing $\tilde{x}=1$, equilibrium beliefs satisfy Bayes rule:

$$\begin{aligned} b_1 &= \Pr(e=1 \mid \tilde{x}=1) \\ &= \frac{\Pr(\tilde{x}=1 \mid e=1)\Pr(e=1)}{\Pr(\tilde{x}=1)} \\ &= \frac{R_0^s E^s}{\Pr(\tilde{x}=1 \mid x=0)\Pr(x=0)+\Pr(\tilde{x}=1 \mid x=1)\Pr(x=1)} \\ &= \frac{R_0^s E^s}{R_0^s(1-\Pr(x=1))+\Pr(x=1)} \end{aligned}$$

Thus, we can write the difference in posteriors as

$$b_1 - b_0 = \frac{E^s(1+\Pr(x=1)R_0^s)}{(\Pr(x=1)-1)(1-R_0^s)}.$$

Recall, $R_0^s < 1$. Because $E^s > 0$, $\Pr(x=1) < 1$. As such, the fraction above is negative. But this means $b_1 < b_0$, a contradiction.

Second, we maintain (ii) and show that a contradiction arises. Consider the agent's reporting decision after outcome $x=1$. Because $R_1^s = 1$, the agent is misclassifying and sending report $\tilde{x}=0$ with probability 1. We have already shown that $\Pr(\tilde{x}=1) \in (0,1)$ implies $R_0^s = 0$, so $\tilde{x}=0$ is being sent with probability 1 after outcome $x=0$. Hence, $\Pr(\tilde{x}=1) = \Pr(\tilde{x}=1 \mid x=1)\Pr(x=1) + \Pr(\tilde{x}=1 \mid x=0)\Pr(x=0) = 0$, a contradiction.

To prove (2), suppose the contrary. That is, there exists equilibrium (s,b) such that $\Pr(\tilde{x}=1) \in (0,1)$ and $C^s = 0$. Because $C^s = 0$, the no crime outcome $x=0$ is always realized, i.e., $\Pr(x=0)=1$. Because $R_0^s = 0$ from result (1), the agent is sending report $\tilde{x}=0$ with probability 1 after $x=0$, i.e., $\Pr(\tilde{x}=1 \mid x=0)=1$. Thus $\Pr(\tilde{x}=1) = \Pr(\tilde{x}=1 \mid x=0)\Pr(x=0)+\Pr(\tilde{x}=1 \mid x=1)\Pr(x=1)=1$, a contradiction.

To prove (3), suppose not. That is, there exists equilibrium (s,b) such that $\Pr(\tilde{x}=1) \in (0,1)$ and $E^s = 1$. Because $E^s = 1$, the no crime outcome $x=0$ is always realized. Because

$R_0^s = 0$ from result (1), the agent is sending report $\tilde{x} = 0$ with probability 1 after the no crime outcome. Like the proof for result (2), this leads to a contradiction. □

A.2 Proof of Proposition 3

We have:

$$\frac{\partial M^s}{\partial R_1^s} = (1 - E^s)C^s > 0.$$

Moreover:

$$\frac{\partial M^s}{\partial C^s} = (1 - E^s)R_1^s + C^s \frac{\partial R_1^s}{\partial C^s} > 0,$$

where the inequality follows from $\frac{\partial R_1^s}{\partial C^s} > 0$, as shown in Lemma 2.

Finally:

$$\begin{aligned}\frac{\partial M^s}{\partial E^s} &= -C^s R_1^s + C^s(1 - E^s)\frac{\partial R_1^s}{\partial E^s} \\ &= C^s \left[-R_1^s + (1 - E^s)\frac{\partial R_1^s}{\partial E^s} \right].\end{aligned}$$

Plugging in $R_1^s = H(\hat{\eta}^*)$ and $\frac{\partial R_1^s}{\partial E^s} = h(\hat{\eta}^*)\frac{\partial \hat{\eta}^*}{\partial E^s}$ yields:

$$\frac{\partial M^s}{\partial E^s} = C^s \left[-H(\hat{\eta}^*) + (1 - E^s)h(\hat{\eta}^*)\frac{\partial \hat{\eta}^*}{\partial E^s} \right]$$

Re-arranging yields the condition stated the main text. Note that by Lemma 2, $\frac{\partial \hat{\eta}^*}{\partial E^s} > 0$. Hence, there are indeed competing effects

A.3 Proof of Proposition 4

Fix the full-support equilibrium (s, b) with effort threshold $\hat{\rho}^*$, i.e., $\Lambda(\hat{\rho}^*) - \hat{\rho}^* = 0$. Consider first how an increase in σ affects the manipulation threshold $\hat{\eta}^*$, which is defined as follows:

$$\bar{\mu}(F(\hat{\rho}^*), G(1 - F(\hat{\rho}^*)), H(\hat{\eta}^* - \sigma)) - \hat{\eta}^* = 0,$$

where $E^s = F(\hat{\rho}^*)$ and $C^s = G(1 - F(\hat{\rho}^*))$. Recall that the PDFs f, g, and h are continuous over their supports, so the CDFs F, G, and H are continuously differentiable (C^1) over their supports. Thus, the implicit function theorem implies that $\hat{\eta}^*$ (when it is interior in the support as in a full-support equilibrium) is C^1 as a function of $\hat{\rho}^*$ and σ. Specifically,

$$\frac{\partial \hat{\eta}^*}{\partial \sigma} = -\frac{\frac{\partial \bar{\mu}}{\partial R_0^s} h(\hat{\eta} - \sigma)(-1)}{\frac{\partial \bar{\mu}}{\partial R_0^s} h(\hat{\eta} - \sigma) - 1} > 0.$$

So an increase in costs increases the threshold, implying higher stakes in Ψ. Also note that $\frac{\partial \hat{\eta}^*}{\partial \sigma} < 1$. Hence, the equilibrium probability of manipulation $R_1^s = H(\hat{\eta}^*(\sigma) - \sigma)$ is decreasing in σ. This proves the first result in Proposition 4.

Second, we argue that Ψ (and as a consequence Λ) is C^1 in ρ and σ. Recall that we can write Ψ as a function of $\hat{\eta}^*$ as follows:

$$\Psi(\hat{\eta}^*) = (1 - H(\hat{\eta}^*))\hat{\eta}^* + \underbrace{\int_{\underline{\eta}}^{\hat{\eta}^*} \eta h(\eta) d\eta}_{\equiv W}.$$

The first expression on the right-hand side of the above equation is C^1 in $\hat{\rho}^*$ and σ because $\hat{\eta}^*$ is C^1. To see that the second expression is also C^1, use Leibniz 's rule to differentiate W to get:

$$\frac{\partial W}{\partial \sigma} = \hat{\eta}^* h(\hat{\eta}^*) \frac{\partial \hat{\eta}^*}{\partial \sigma} \qquad \text{and} \qquad \frac{\partial W}{\partial \hat{\rho}^*} = \hat{\eta}^* h(\hat{\eta}^*) \frac{\partial \hat{\eta}^*}{\partial \hat{\rho}^*}.$$

Because $\hat{\eta}^*$ is C^1 and h is continuous, these derivatives are also continuous. So Ψ is C^1 in $\hat{\rho}^*$ and σ. Thus, Λ is also C^1 given its definition and the C^1 properties of G, F, and Ψ.

Third, we focus on how $\hat{\rho}^*$ changes as a function of σ. By the implicit function theorem, we have:

$$\frac{\partial \hat{\rho}^*}{\partial \sigma} = -\frac{\frac{\partial \Lambda}{\partial \sigma}}{\frac{\partial \Lambda}{\partial \hat{\rho}} - 1},$$

which is well-defined because $\Lambda'(\hat{\rho}^*) \neq 1$. Since σ affects Λ only through Ψ, it is enough to investigate how Ψ changes as σ changes. We have:

$$\begin{aligned} \frac{d\Psi}{d\sigma} = & -h(\hat{\eta}^* - \sigma)\left(\frac{\partial \hat{\eta}^*}{\partial \sigma} - 1\right)\hat{\eta}^* + (1 - H(\hat{\eta}^* - \sigma))\frac{\partial \hat{\eta}^*}{\partial \sigma} + \\ & \hat{\eta}^* h(\hat{\eta}^* - \sigma)\frac{\partial \hat{\eta}^*}{\partial \sigma} - \left(\underline{\eta} + \sigma\right) h(\underline{\eta}) + \int_{\underline{\eta}+\sigma}^{\hat{\eta}^*} \eta h'(\eta - \sigma)(-1) d\eta \\ & = h(\hat{\eta}^* - \sigma)\hat{\eta}^* - \left(\underline{\eta} + \sigma\right) h(\underline{\eta}) + (1 - H(\hat{\eta}^* - \sigma))\frac{\partial \hat{\eta}^*}{\partial \sigma} - \int_{\underline{\eta}+\sigma}^{\hat{\eta}^*} \eta h'(\eta - \sigma) d\eta \end{aligned}$$

This cannot be signed in general, but it is in particular positive if H is the Uniform distribution. The reason is that in this case, the derivative of the density is 0 and the density is a constant. Since $\hat{\eta}^* > \underline{\eta} + \sigma$ in an interior equilibrium, we have that $\frac{\partial \Psi}{\partial \sigma} > 0$.

Thus, we need to verify that $\frac{\partial \Lambda}{\partial \hat{\rho}}\big|_{\hat{\rho}=\hat{\rho}^*} < 1$. We show that this is implied by the game having a unique equilibrium, i.e., $\Lambda(\hat{\rho}) = \hat{\rho}$ has a unique solution. We also assume that $G(1) < 1$, which makes b_0 well-defined for all E^s, $F(G(1)\beta) > 0$, and $F(G(0)(\beta + 1)) < 1$.

To cover both the case in which F has bounded support and in which the support of F is the entire real line, we denote by $\underline{\rho}$ the minimum of the support of F or $\underline{\rho} = -\infty$ in the case that the support is not bounded below. Similarly, $\overline{\rho}$ is the maximum of the support of

F or $\overline{\rho} = \infty$ in the support is not bounded above.

First observe that, as $\hat{\rho} \to \underline{\rho}$, $F(\hat{\rho}) \to 0$. Because $G(1) < 1$, $\bar{\mu}(0, G(1), R_1^s) = 0$ for all $R_1^s \in [0,1]$. As such, $\hat{\eta} = 0$ is the unique solution to $\bar{\mu}(0, G(1), H(\hat{\eta})) = \hat{\eta}$, which means $\Psi(0, G(1)) = 0$. Because Ψ is continuous,

$$\lim_{\hat{\rho} \to \underline{\rho}} \Psi(F(\hat{\rho}), G(1 - F(\hat{\rho}))) = \Psi(0, G(1)) = 0.$$

Putting this altogether, gives us

$$\begin{aligned}\lim_{\hat{\rho} \to \underline{\rho}} \Lambda(\hat{\rho}) &= \lim_{\hat{\rho} \to \underline{\rho}} G(1 - F(\hat{\rho}))\beta + \lim_{\hat{\rho} \to \underline{\rho}} G(1 - F(\hat{\rho}))\Psi(F(\hat{\rho}), G(1 - F(\hat{\rho}))) \\ &= G(1)\beta + G(1)0 \\ &= G(1)\beta > \underline{\rho}.\end{aligned}$$

The last inequality follows from $F(G(1)\beta) > 0$, which implies $G(1)\beta > \underline{\rho}$ because $F(\underline{\rho}) = 0$ and F is increasing.

Similarly, consider $\hat{\rho} \to \overline{\rho}$, which means $F(\hat{\rho}) \to 1$. Then $\hat{\eta}^* = 1$ is the unique solution to $\bar{\mu}(1, G(0), H(\hat{\eta})) = \hat{\eta}$, which means $R_1^s = H(1)$. Because Ψ is continuous,

$$\lim_{\hat{\rho} \to \overline{\rho}} \Psi(F(\hat{\rho}), G(1 - F(\hat{\rho}))) = \Psi(1, G(0)).$$

As a result:

$$\lim_{\hat{\rho} \to \overline{\rho}} \Lambda(\hat{\rho}) = G(0)\left(\beta + \Psi(1, G(0))\right) \leq G(0)\left(\beta + 1\right).$$

The inequality follows because Ψ is always bounded above by 1. We require $G(0)\left(\beta + 1\right) < \overline{\rho}$, which is implied by $F(G(0)(\beta + 1)) < 1$.

Now consider $Q(\hat{\rho}) \equiv \Lambda(\hat{\rho}) - \hat{\rho}$. By above, Q is continuously differentiable as a function of ρ, and

$$\lim_{\hat{\rho} \to \underline{\rho}} Q(\hat{\rho}) > 0$$

while

$$\lim_{\hat{\rho} \to \overline{\rho}} Q(\hat{\rho}) < 0.$$

If there is a unique solution such that $Q(\hat{\rho}^*) = 0$, it must be the case that $Q'(\hat{\rho}^*) < 0$. To see this, suppose not. Then $Q'(\hat{\rho}^*) \geq 0$. But then $\Lambda'(\hat{\rho}^*) \neq 1$ implies $Q'(\hat{\rho}^*) > 0$. Because Q is C^1, there exist $\epsilon > 0$ such that $Q(\hat{\rho}^* + \epsilon) > 0$. Recall that $\lim_{\hat{\rho} \to \overline{\rho}} Q(\hat{\rho}) < 0$. Because Q is continuous, the intermediate value theorem implies there exists $\hat{\rho}^{**} \in (\hat{\rho}^*, \bar{\rho})$ such that $Q(\hat{\rho}^{**}) = 0$. But then $\hat{\rho}^{**}$ pins down another full-support equilibrium, a contradition. Thus, $Q'(\hat{\rho}^*) < 0$, which means that $\frac{\partial \Lambda}{\partial \hat{\rho}}\big|_{\hat{\rho} = \hat{\rho}^*} < 1$.

A.4 Proof of Proposition 5

Note that Lemma 1 establishes that if the actual law enforcement statistic is produced by $x = (1 - e)c$, and both reports are sent in equilibrium, then $R_0^s = 0$ and $R_1^s < 1$. Then M^s takes the form $M^s = (1 - E^s)C^s R_1^s \geq 0$, as discussed in Section 4.1. To establish the corresponding result when the actual law enforcement statistic is produced by $x = ec$, we first prove the following lemma:

Lemma A.1. *Assume $x = ec$. If (s, b) is a equilibrium such that both reports are sent, i.e.,* $\Pr(\tilde{x} = 1) \in (0, 1)$*, then the following hold:*

1. *$R_1^s = 0$ and $R_0^s < 1$.*

2. *either $C^s < 1$ or $E^s < 1$.*

Proof. Suppose the contrary. That is, there exists equilibrium (s, μ) is a equilibrium such that $\Pr(\tilde{x} = 1) \in (0, 1)$ and either (i) $R_1^s > 0$ or (ii) $R_0^s = 1$.

First, we maintain (i) and show that a contradiction arises. If $R_1^s > 0$, then since $\eta > 0$, we must have $b_0 > b_1$. Now consider the reporting decision after $x = 0$ (may be off-the-path). The expected utility of reporting $\tilde{x} = 1$ is $b_1 - \eta$ while the expected utility of reporting $\tilde{x} = 0$ is b_0. Thus $b_0 > b_1 - \eta$ and $R_0^s = 0$.

Consider several cases:

1. If $\Pr(x = 0) = 1$, $\Pr(\tilde{x} = 1) = 0$, a contradiction.
2. If $\Pr(x = 0) < 1$ and $R_1^s = 1$, then again $\Pr(\tilde{x} = 1) = 0$, a contradiction.
3. If $\Pr(x = 0) < 1$ (which in particular implies $C^s > 0$) and $R_1^s < 1$, then both b_1 and b_0 can be computed from Bayes' rule:

$$b_1 = \frac{E^s \left[C^s(1 - R_1^s) + (1 - C^s)0\right]}{E^s C^s(1 - R_1^s) + (1 - E^s C^s)0} = \frac{E^s C^s(1 - R_1^s)}{E^s C^s(1 - R_1^s)}$$

Also:

$$b_0 = \frac{E^s \left[C^s R_1^s + (1 - C^s)1\right]}{E^s C^s R_1^s + (1 - E^s C^s)1}$$

By inspection, $b_1 = 1$ while $b_0 \leq 1$ (with strict inequality if $E^s < 1$), a contradiction.

Second, we maintain (ii) and show that a contradiction arises. Consider the agent's reporting decision after outcome $x = 0$. Because $R_0^s = 1$, the agent is misclassifying and sending report $\tilde{x} = 1$ with probability 1. We have already shown that $\Pr(\tilde{x} = 1) \in (0, 1)$ implies $R_1^s = 0$, so $\tilde{x} = 1$ is being sent with probability 1 after outcome $x = 1$. Hence, $\Pr(\tilde{x} = 0) = \Pr(\tilde{x} = 0|x = 1)\Pr(x = 1) + \Pr(\tilde{x} = 0|x = 0)\Pr(x = 0) = 0$, a contradiction.

To prove (2), suppose the contrary. That is, there exists equilibrium (s, b) such that $\Pr(\tilde{x} = 1) \in (0, 1)$ and $C^s = E^s = 1$. Because $C^s = 1 = E^1$, the crime outcome $x = 1$ is always realized, i.e., $Pr(x = 1) = 1$. Because $R_1^s = 0$ from result (1), the agent is sending report $\tilde{x} = 1$ with probability 1 after $x = 1$, i.e., $Pr(\tilde{x} = 1|x = 1) = 1$. Thus $Pr(\tilde{x} = 1) = Pr(\tilde{x} = 1|x = 0)Pr(x = 0) + Pr(\tilde{x} = 1|x = 1)Pr(x = 1) = 1$, a contradiction. □

Proposition 5 then follows because the lemma allows us to write measurement error as:

$$M^s = \Pr(x = 1|s) - \Pr(\tilde{x} = 1|s) = E^s C^s - [E^s C^s + (1 - E^s C^s)R_0^s] = -(1 - E^s C^s)R_0^s \leq 0$$

A.5 Proof of Proposition 6

This follows from:

$$\frac{\partial \Pr(x = 1)}{\partial \theta} - \frac{\partial \Pr(\tilde{x} = 1)}{\partial \theta} = \frac{\partial \Pr(x = 1) - \Pr(\tilde{x} = 1)}{\partial \theta} = \frac{\partial M^s}{\partial \theta}.$$

B Sufficient conditions for all equilibria to have full support

Fact B.1. *In every equilibrium (s, b), $C^s \in [G(0), G(1)]$.*

Lemma B.1 states three sufficient conditions for the reporting subgame to be well defined in the sense that $E^s \in (0, 1)$ means there is uncertainty over whether or not the agent exerted effort and $C^s > 0$, along with $E \in (0, 1)$, means that both LE statistics $x \in \{0, 1\}$, i.e., signals about effort, arise with positive probability.

Lemma B.1. *The following implications hold:*

1. *$G(1 - F(0)) > 0$ implies $C^s > 0$ in equilibrium (s, b).*

2. *$\underline{\eta} = 0$ and $0 < F(G(1)\beta)$ imply $E^s > 0$ in equilibrium every equilibrium (s, b).*

3. *$F(G(0)\beta + 1) < 1$ implies $E^s < 1$ in equilibrium (s, b).*

Proof. To see (1), suppose the contrary. So there exists equilibrium (s, b) such that $C^s = 0$. Then $x = 0$, regardless of whether the agent works or shirks. So we can write the expected utility of working as $-\rho + \int \max\{b_0, b_1 - \eta\}h(\eta)d\eta$. The expected utility of not working is $\int \max\{b_0, b_1 - \eta\}h(\eta)d\eta$. As such, the agent works if and only if $\rho \leq 0$, so $E^s = F(0)$. If the target engages in illegal activity, their expected payoff is $(1 - E^s)1 - \gamma$. If the target does not engage, their payoff is 0. So $C^s = G(1 - E^s) = G(1 - F(0)) > 0$.

To see (2), suppose the contrary. So there exists equilibrium (s, b) such that $E^s = 0$ when $0 < F(G(1)\beta)$. First, either $\Pr(\tilde{x} = 1) > 0$ or $\Pr(\tilde{x} = 0) > 0$. If $\Pr(\tilde{x} = 1) > 0$, then Bayes rules implies $b_1 = 0$. If $b_0 > 0$, then agents in the reporting subgame with $\eta < b_0 - b_1$ would submit $\tilde{x} = 0$. Because $\min \operatorname{supp}(H) = \underline{\eta} = 0$, $H(b_0 - b_1) > 0$. Thus, report $\tilde{x} = 0$ would be sent with positive probability given s, implying $b_0 = 0$. A similar argument shows that $\Pr(\tilde{x} = 0) > 0$ implies $b_0 = b_1 = 0$. Second, $E^s = 0$, so the target chooses $c = 1$ if and essentially only if $\gamma < 1$. Thus, $C^s = G(1)$. Finally, consider the agent with $\rho' < G(1)\beta$. If she exerts effort, then her expected payoff is $C^s\beta - \rho$. If she does not exert effort, then her expected payoff is 0. Because $\rho' < G(1)\beta$, this agent works. Thus $E^s \geq F(G(1)\beta) > 0$.

To see (3), suppose the contrary. So there exists equilibrium (s, b) such that $E^s = 1$ when $F(G(0)\beta + 1) < 1$. There exists report $r \in \{0, 1\}$ such that $\Pr(\tilde{x} = r) > 0$. After such a report, Bayes rule implies $b_r = 1$. First, $E^s = 1$, so the target chooses $c = 1$ if and essentially only if $\gamma < 0$. Thus, $C^s = G(0)$. Consider the agent with $\rho' > G(0)\beta + 1$. If this agent works, at best she expects to get $C^s\beta - \rho' + 1$, where $C^s\beta - \rho'$ is their expected utility from the encounter after working, and 1 is their maximum payoff from the reporting and assessment stage. If this agent does not work, at worst she expects to get 0, where 0 is their payoff in the encounter after not working and 0 is the minimum payoff from the reporting and assessment stage. Because $\rho' > G(0)\beta + 1$, this agent shirks. Thus, $E^s \leq 1 - F(G(0)\beta + 1) < 1$. □

Lemma B.2. *If $H(F(G(1)(\beta + \mathbb{E}[\eta]))) < 1$, then $R_0^s = 0$ and $R_1^s < 1$ in every equilibrium (s, b).*

Proof. To see that $R_0^s = 0$, suppose the contrary. That is, suppose there exists equilibrium (s, b) such that $R_0^s > 0$. Thus, some agents with lying costs η are reclassifying after outcome $x = 0$. After these agents reclassify by reporting $\tilde{x} = 1$, their (net-of-encounter) payoffs are $b_1 - \eta$. If these agents were not to reclassify by reporting $\tilde{x} = 0$, there (net-of-encounter) expected payoffs are b_0. Because $\underline{\eta} \geq 0$ and $\Pr(\eta > \underline{\eta}) = 1$, $b_1 > b_0$. Furthermore, $b_1 > b_0$ and $\eta \geq 0$ imply that agents with outcome $x = 1$ will never reclassify, so $R_1^s = 0$.

Now we can write probability that report $\tilde{x} = 0$ is sent in equilibrium as

$$\begin{aligned}
\Pr(\tilde{x} = 0) &= \Pr(\tilde{x} = 0 | x = 1) \Pr(x = 1) + \Pr(\tilde{x} = 0 | x = 0) \Pr(x = 0) \\
&= R_1^s \Pr(x = 1) + (1 - R_0^s) \Pr(x = 0) \\
&= (1 - R_0^s)(E^s + (1 - E^s)(1 - C^s)).
\end{aligned}$$

Above, the first equality is the law of total expectations, and the second follows from the definition of R_x^s. The third follows because $R_1^s = 0$. We consider two cases.

Case 1: $\Pr(\tilde{x} = 0) > 0$. That is, the report $\tilde{x} = 0$ is being sent with positive probability in equilibrium (s, b). After seeing $\tilde{x} = 0$, equilibrium beliefs are then computed via Bayes rule:

$$\begin{aligned}
b_0 &= \Pr(e = 1 | \tilde{x} = 0) \\
&= \frac{\Pr(\tilde{x} = 0 | e = 1) \Pr(e = 1)}{\Pr(\tilde{x} = 0)} \\
&= \frac{(1 - R_0^s) E^s}{(E^s + (1 - E^s)(1 - C^s))(1 - R_0^s)}.
\end{aligned}$$

Second, note that

$$\Pr(\tilde{x}=1) = (E^s + (1-E^s)(1-C^s))R_0^s + (1-E^s)C^s > 0.$$

Thus, after seeing $\tilde{x}=1$, equilibrium beliefs are also computed via Bayes rule:

$$\begin{aligned} b_1 &= \Pr(e=1|\tilde{x}=1) \\ &= \frac{\Pr(\tilde{x}=1|e=1)\Pr(e=1)}{\Pr(\tilde{x}=1)} \\ &= \frac{R_0^s E^s}{(E^s + (1-E^s)(1-C^s))R_0^s + (1-E^s)C^s}. \end{aligned}$$

After some algebra, we can compute the difference in beliefs as

$$b_1 - b_0 = \frac{C^s(E^s-1)E^s}{(1-C^s(1-E^s))(C^s(1-R_0^s)(1-E^s)+R_0^s)} \leq 0,$$

because $E^s \leq 1$. This contradicts $b_1 > b_0$, however.

Case 2: $\Pr(\tilde{x}=0)=0$. That is, the report $\tilde{x}=1$ is being sent with probability 1 in equilibrium (s,b). As such $b_1 = E^s$, that is the posterior should equal the prior. Recall, $b_1 > b_0$, so $E^s > 0$, which in turn implies $\Pr(x=0) > 0$. Because $\Pr(\tilde{x}=0)=0$, after outcome $x=0$, the agent must surely misreport, which means $b_1 - \eta \geq b_0$ for all $\eta \in \operatorname{supp} H$. As such $H(b_1-b_0)=1$, and we can substitute for $H(E^s-b_0)=1$ as $b_1=E^s$. Then note that H is increasing and $b_0 \geq 0$. So $H(E^s-b_0)=1$ implies $H(E^s)=1$. Next, we derive an upper bound on E^s and show that the equality, $H(E^s)=1$, cannot hold. After working in the encounter stage, the crime outcome will be $x=0$, which will be reclassified as $R_0^s=1$. Thus, the agent's expected utility from working is

$$\beta C^s - \rho + b_1 - \mathbb{E}[\eta].$$

After not working, with probability C^s, the outcome will be $x=1$, which will not be reclassified, but with probability $1-C^s$ the outcome will be $x=0$, which will be reclassified. Thus, the agent's expected utility from not working is

$$(1-C^s)(b_1-\mathbb{E}[\eta]) + C^s b_1.$$

Comparing these utilities reveals that, if $\rho > C^s(\beta - \mathbb{E}[\eta])$, then the agent will not work. Thus, E^s is bounded above by $F(C^s(\beta-\mathbb{E}[\eta]))$. Because F is increasing and $C^s \geq 0$, E^s is bounded above by $F(C^s(\beta+\mathbb{E}[\eta]))$. Because $C^s \in [G(0), G(1)]$, E^s is bounded above by $F(G(1)(\beta+\mathbb{E}[\eta]))$. Thus, $H(E^s) \leq H(F(G(1)(\beta+\mathbb{E}[\eta]))) < 1$. But then $H(E^s) \neq 1$, which is the desired contradiction.

Hence, we conclude $R_0^s = 0$. Next we argue that $R_1^s < 1$ in every equilibrium (s,b). To see this, suppose not. That is, suppose there exists equilibrium (s, b) such that $R_1^s = 1$. By previous argument, we know $R_0^s = 0$. Thus, report $\tilde{x} = 0$ is being sent with probability 1, so $b_0 = \Pr(e = 1|\tilde{x} = 0) = E^s$. Furthermore, for all $\eta \in \text{supp } H$, we must have $b_0 - \eta \geq b_1$, or $E^s - \eta \geq b_1$. Thus, $R_1^s = H(E^s - b_1) = 1$. As above, a necessary condition for this equality to hold is $H(E^s) = 1$. Now we derive an upper bound on E^s, and show that the equality, $H(E^s) = 1$, cannot hold. If the agent works, then her payoff is $\beta C^s - \rho + b_0$. If the agent shirks, then her payoff is $b_0 - C^s\mathbb{E}[\eta]$. Comparing these utilities reveals that the Agent does not work if $\rho > C^s(\beta + \mathbb{E}[\eta])$, so E^s is bounded above by $F(C^s(\beta + \mathbb{E}[\eta]))$. This expression is strictly increasing in C^s, where $C^s \leq G(1)$. So E^s is bounded above by $F(G(1)(\beta + \mathbb{E}[\eta]))$. Thus, $H(E^s) \leq H(F(G(1)(\beta + \mathbb{E}[\eta]))) < 1$, a contradiction. □

The final lemma says that when both law enforcement outcomes $x \in \{0, 1\}$ are realized, then a truthful equilibrium does not exist if $\underline{\eta} = 0$.

Lemma B.3. *Consider an equilibrium (s, b) such that $E^s < 1$, $C^s > 0$, $R_0^s = 0$. If $\underline{\eta} = 0$, then $R_1^s > 0$*

Proof. Suppose the contrary. That is, assume $\underline{\eta} = 0$, and suppose (s, b) is an equilibrium such that $E^s < 1$, $C^s > 0$, and $R_0^s = R_1^s = 0$. First notice that $E^s < 1$ and $C^s > 0$ imply both LE outcomes $x \in \{0, 1\}$ occur with positive probability. Second, because the agent is truthful and is not reclassifying, both reports $\tilde{x} \in \{0, 1\}$ are sent with positive probability.

After report $\tilde{x} = 1$, the agent is not reclassifying (as $R_1^s = 0$) so we must have $b_1 = 0$ by Bayes rule. After report $\tilde{x} = 0$, the agent is not reclassifying, and b_0 can be computed using Bayes rule as well:

$$\begin{aligned} b_0 &= \Pr(e = 1 \mid \tilde{x} = 0) \\ &= \frac{\Pr(\tilde{x} = 0 \mid e = 1)\Pr(e = 1)}{\Pr(\tilde{x} = 0 \mid e = 1)\Pr(e = 1) + \Pr(\tilde{x} = 0 \mid e = 0)\Pr(e = 0)} \\ &= \frac{E^s}{E^s + (1 - E^s)(1 - C^s)} > 0. \end{aligned}$$

But then $b_0 > b_1$. Thus, in the reporting subgame after outcome $x = 1$, agents with $\eta < b_0 - b_1$ send message $\tilde{x} = 0$. Hence, $R_1^s \geq H(b_0 - b_1)$. Because $\underline{\eta} = \min \text{supp}\, H = 0$, $H(b_0 - b_1) > 0$, so $R_1^s > 0$. This established the desired contradiction. □

Definition B.1. *Equilibrium (s, b) has full support if $E^s \in (0, 1)$, $C^s > 0$, $R_0^s = 0$ and $R_1^s \in (0, 1)$*

Proposition B.1. *All equilibria have full support if the following conditions hold:*

1. *$G(1 - F(0)) > 0$,*

2. *$0 < F(G(1)\beta)$,*

3. $F(G(0)\beta + 1) < 1$,

4. $\underline{\eta} = 0$,

5. $H(F(G(1)(\beta + \mathbb{E}[\eta]))) < 1$.

Proof. The result follows immediately from Lemmas B.1, B.2, and B.3. □

The conditions in Proposition B.1 are fairly innocuous. Conditions 1–3 are satisfied if F and G have full support over $\mathbb{R}$, e.g., F and G are normal distributions. If F and G have interval support, e.g., $\operatorname{supp} F = [\underline{\rho}, \bar{\rho}]$ and $\operatorname{supp} G = [\underline{\gamma}, \bar{\gamma}]$, then Conditions 1–3 are satisfied if $\underline{\rho} = \underline{\gamma} = 0$ and $\bar{\rho} > 1$. Conditions 4 and 5 are satisfied if H is the exponential distribution or if it is the uniform distribution with over $[0, 1]$ with $F(G(1)(\beta + 0.5)) < 1$.

C Agents with types and relevant reputation concerns

In the baseline version of the model, the agent internalizes the beliefs of a third party about whether or not she exerted effort. Although this could be interpreted as retrospective support or capturing how an agency's funding depends on third parties believing that they are hard working, it could be the case that agent cares about maintain a reputation that she is a hard-working innate type. In this Appendix, we consider such a version of the model.

In this version, there are two agent types: a diligent type ($\tau = D$) and a lazy type ($\tau = L$). Lazy types of agents always have high costs of effort, i.e., $\rho = \infty$. That is, lazy types always choose $e = 0$. Diligent types of agents have effort costs ρ drawn from F in the baseline model. The timing and interaction are the same as in the baseline model, but now the third party beliefs are

$$b_{\tilde{x}} = \Pr(\tau = D \mid \tilde{x}).$$

Specifically, the following summarize the timeline of this game:

1. Nature draws the type $\tau \in \{D, L\}$, with $\Pr(\tau = D) = \pi$.
2. T observes the opportunity cost $\gamma \sim G$. A observes τ and the cost of effort ρ, where $\rho \sim F$ if $\tau = D$ and $\rho = \infty$ if $\tau = L$.
3. Simultaneously, A chooses effort e and T chooses behavior c.
4. Enforcement payoffs are realized.
5. The law enforcement outcome is produced with the technology $x = (1 - e)c$.
6. A sees x and cost of manipulating data $\eta \sim H$.
7. A writes a report $\tilde{x} \in \{0, 1\}$.
8. A receives reporting payoffs $b_{\tilde{x}} - \eta\mathbb{I}[\tilde{x} \neq x]$, where $b_{\tilde{x}} = \Pr(\tau = D \mid \tilde{x})$.

C.1 Analysis

Notice that conditional on the realizations of ρ and η, the expected payoffs for the two types of actors are identical. As such, strategies and equilibria are the same as in the baseline model. In addition, it is useful to work with the corresponding higher-order probabilities:

$$E^s_\tau = \Pr(e = 1 \mid s, \tau) = \begin{cases} \int \mathbb{I}[s^{\texttt{en}}_A(\rho) = 1] f(\rho) d\rho & \text{if } \tau = D \\ 0 & \text{if } \tau = L \end{cases},$$

where E^s_τ is the probability that A exerts high effort given type τ, where we are explicitly imposing the equilibrium condition that L-types with $\rho = \infty$ will never exert high effort. In addition, consider agent A with type $\tau \in \{D, L\}$ in the reporting subgame with data manipulation costs η. When she sends a report $\tilde{x}$, she receives $b_{\tilde{x}} - \eta\mathbb{I}[\tilde{x} \neq x]$. Notice, this payoff is independent of type τ. Thus, A with type τ and cost η sends report $\tilde{x}$ after outcome x if and only if A with type $\tau' \neq \tau$ and cost η sends report $\tilde{x}$ after outcome x. So conditioning on η and x, the decision to reclassify does not depend on type τ. In addition, η is a random variable drawn from H, which also does not depend on τ. As such, the agents of $\tau = D$ and $\tau = L$ will reclassify at identical rates, so we define:

$$R^s_x = \Pr(\tilde{x} \neq x \mid x, s) = \int \mathbb{I}[s^{\texttt{re}}_A(x, \eta) \neq x] h(\eta) d\eta.$$

As in the baseline model, we want to characterize full-support equilibria. In this context, we study equilibria (s, b) such that $E^s_D \in (0, 1)$, $C^s > 0$, $R^s_0 = 0$, and $R^s_1 \in (0, 1)$.

First consider the reporting stage. After a message $\tilde{x} = 1$, we want to compute

$$\begin{aligned} b_1 &= \Pr(D \mid \tilde{x} = 1) \\ &= \frac{\Pr(\tilde{x} = 1 \mid \tau = D) \Pr(\tau = D)}{\Pr(\tilde{x} = 1)} \\ &= \frac{(1 - E^s_D) C^s (1 - R^s_1) \pi}{\Pr(\tau = D) \Pr(\tilde{x} = 1 \mid \tau = D) + \Pr(\tau = L) \Pr(\tilde{x} = 1 \mid \tau = L)} \\ &= \frac{(1 - E^s_D) C^s (1 - R^s_1) \pi}{\pi (1 - E^s_D) C^s (1 - R^s_1) + (1 - \pi)(1 - E^s_L) C^s (1 - R^s_1)} \\ &= \frac{(1 - E^s_D)(1 - R^s_1) \pi}{\pi (1 - E^s_D)(1 - R^s_1) + (1 - \pi)(1 - R^s_1)} \end{aligned}$$

After message $\tilde{x} = 0$, we want to compute

$$\begin{aligned}
b_0 &= \frac{\Pr(\tilde{x} = 0 \mid \tau = D)\Pr(\tau = D)}{\Pr(\tilde{x} = 0)} \\
&= \frac{(1 - \Pr(\tilde{x} = 1 \mid \tau = D))\Pr(\tau = D)}{1 - \Pr(\tilde{x} = 1)} \\
&= \frac{[1 - (1 - E^s_D)C^s(1 - R^s_1)]\pi}{1 - [\Pr(\tau = D)\Pr(\tilde{x} = 1 \mid \tau = D) + \Pr(\tau = L)\Pr(\tilde{x} = 1 \mid \tau = L)]} \\
&= \frac{[1 - (1 - E^s_D)C^s(1 - R^s_1)]\pi}{1 - [\pi(1 - E^s_D)C(1 - R^s_1) + (1 - \pi)(1 - E^s_L)C^s(1 - R^s_1)]} \\
&= \frac{[1 - (1 - E^s_D)C(1 - R^s_1)]\pi}{1 - C^s[\pi(1 - E^s_D)(1 - R^s_1) + (1 - \pi)(1 - R^s_1)]}.
\end{aligned}$$

Finally, we can compute the difference in posteriors as

$$b_0 - b_1 = \frac{E^s_D\pi(1 - \pi)}{(1 - E^s_D\pi)(1 - C^s(1 - E^s_D\pi)(1 - R^s_1))} \equiv \bar{\mu}(E^s_D, C^s, R^s_1) > 0.$$

The next result follows from the above derivation of $\bar{\mu}$.

Proposition C.1. *When the agents have types and reputation concerns about types, the difference in posteriors $\bar{\mu}$ has partial derivatives with signs that match the baseline model, i.e., $\frac{\partial\bar{\mu}}{\partial E^s_D} > 0$, $\frac{\partial\bar{\mu}}{\partial C^s} > 0$, and $\frac{\partial\bar{\mu}}{\partial R^s_1} < 0$.*

To finish characterizing full-support equilibria, regardless of τ, after crime outcome $x = 1$, the agent compares b_1 to $b_0 - \eta$. Thus, a threshold strategy is optimal. Because $\eta \sim H$, for all candidates, we can write the equilibrium threshold $\hat{\eta}$ as solving the following equation:

$$\hat{\eta} = \bar{\mu}(E^s_D, C^s, H(\hat{\eta})). \tag{C.1}$$

Equation C.1 matches Equation 1 from the baseline model, so the next result follows from an identical application of the implicit function theorem.

Proposition C.2. *When the agents have types and reputation concerns about types, the equilibrium threshold $\hat{\eta}^*$ has partial derivatives with signs that match the baseline model, i.e.,*

$$\frac{\partial\hat{\eta}^*}{\partial E^s_D} = -\frac{\frac{\partial\bar{\mu}}{\partial E^s_D}}{\frac{\partial\bar{\mu}}{\partial R^s_1}h(\hat{\eta}^*) - 1} > 0 \qquad \textit{and} \qquad \frac{\partial\hat{\eta}^*}{\partial C^s} = -\frac{\frac{\partial\bar{\mu}}{\partial C^s}}{\frac{\partial\bar{\mu}}{\partial R^s_1}h(\hat{\eta}^*) - 1} > 0.$$

Thus, we need to solve for E^s_D and C^s. If T engages in illicit activity, then his expected payoff is $(1 - \pi E^s_D) - \gamma$. If not, his expected payoff is 0. So a threshold strategy is also optimal where

$$\hat{\gamma} = (1 - \pi E^s_D) \tag{C.2}$$

and $C^s = G(1 - \pi E_D^s)$.

For A with $\tau = D$, if she exerts effort, then the crime outcome will for sure be $x = 0$ and the report will be $\tilde{x} = 0$. Thus, after exerting effort, her expected payoff is $C^s\beta - \rho + b_0$. If A with $\tau = D$ does not exert effort, then her expected payoff is

$$(1 - C^s)b_0 + C^s \left[(1 - R_1^s)b_1 + R_1^s(b_0 + \mathbb{E}[\eta | \eta \leq \bar{\mu}(E_D^s, C^s, R_1^s)])\right].$$

Thus, the diligent agent uses a threshold strategy, with threshold $\hat{\rho}$. So the diligent agent exerts effort if and only if $\rho < C^s \left[\beta + \Psi(E_D^s, C^s)\right]$ where

$$\begin{aligned}\Psi(E_D^s, C^s) &= (1 - R_1^s)\bar{\mu}(E_D^s, C^s, R_1^s) + R_1^s\mathbb{E}[\eta|\eta \leq \bar{\mu}(E_D^s, C^s, R_1^s)] \\ &= (1 - H(\hat{\eta}^*))\hat{\eta}^* + \int_{\underline{\eta}}^{\hat{\eta}^*} \eta h(\eta) d\eta.\end{aligned}$$

Notice that the derivation of Ψ matches the derivation in the baseline model. As such, the next result follows immediately from the previous two propositions in this section.

Proposition C.3. *When the agents have types and reputation concerns about types, the equilibrium dynamic incentives to work, Ψ, have partial derivatives with signs that match the baseline model, i.e., $\frac{\partial \Psi}{\partial E_D^s} > 0$ and $\frac{\partial \Psi}{\partial C^s} > 0$.*

Recall that $E_D^s = F(\hat{\rho})$ and the target's best response to $\hat{\rho}$ is $C^s = G(1 - \pi F(\hat{\rho}))$. Plugging this into the preceding expressions, a full-support equilibrium (s, b) is characterized by a threshold strategy, $\hat{\rho}^*$, that solves

$$\underbrace{G(1 - \pi F(\hat{\rho}))[\beta + \Psi(F(\hat{\rho}), G(1 - \pi F(\hat{\rho})))]}_{\equiv \Lambda(\hat{\rho})} = \hat{\rho}. \tag{C.3}$$

Together, Equations C.1, C.2, and C.3 characterize full-support equilibria in an identical manner as in the statement of Proposition 1.

C.2 Comparison to baseline model

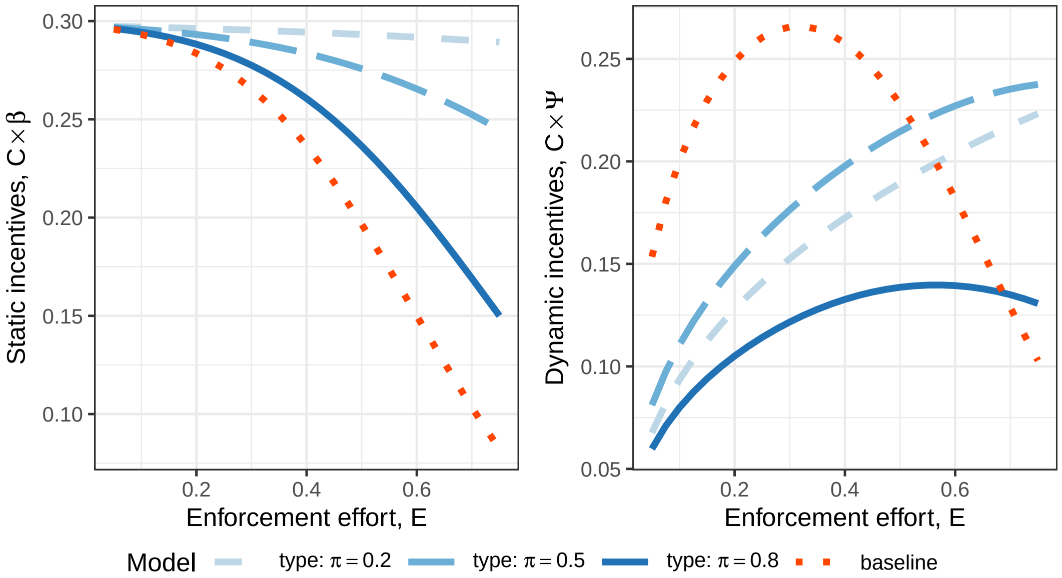
To compare the two models, we consider a numerical example where we assume the following:

$$\gamma \sim \mathcal{N}(0.4, 0.25) \qquad \rho \sim \mathcal{N}(0.5, 0.2) \qquad \eta \sim \mathcal{U}(0, 0.75) \qquad \beta = 0.3.$$

These distributional assumptions are fixed across the baseline and type-extension model. For the latter, we treat the prior probability of a diligent type π as a free parameter.

To begin the model comparison, Figure C.1 graphs an agent's incentives (vertical axis) for effort given a fixed belief about effort (horizontal axis). These are partial equilibrium

Figure C.1: Incentives for enforcement effort across the models.

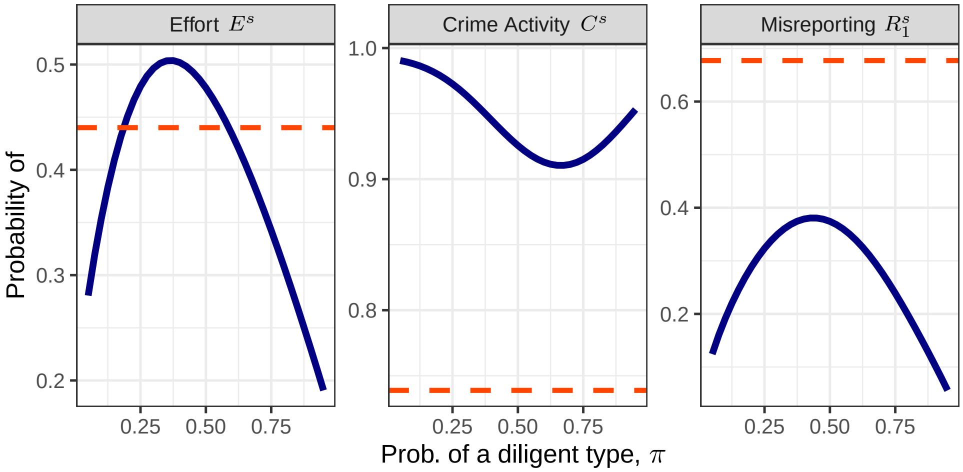


Notes: Horizontal axis is the fixed probability of effort which is E^s in the baseline model and E^s_D in the type-extension model. Crime rates are computed assuming equilibrium behavior of the target, i.e., $C^s = G(1 - E^s)$ in the baseline and $C^s = G(1 - \pi E^s_D)$ in the type-extension model. Example generated assuming $\gamma \sim \mathcal{N}(0.4, 0.25)$, $\beta = 0.3$, $\rho \sim \mathcal{N}(0.5, 0.2)$, and $\eta \sim \mathcal{U}(0, 0.75)$. The prior of the diligent type, $\pi \in (0, 1)$ is a free parameter for the type-extension model.

results, so the horizontal axis means a fixed E^s in the baseline model or a fixed E^s_D in the type-extension model. The left panel shows the static incentives for effort. Notice, the baseline model has the smallest static incentives. Recall that in the baseline model $C^s = G(1 - E^s)$ in Equation 2, but in the type extension $C^s = G(1 - \pi E^s_D)$ in Equation C.2. Thus, there is greater criminal activity in the type extension because with probability $1 - \pi > 0$, an agent is lazy and never exerts effort. As a consequence, criminal activity and therefore static incentives are larger as the prior probability of a diligent type decreases.

Figure C.1's right panel graphs the dynamic incentives for effort, and there are two key takeaways. Most importantly, the baseline model generally has larger dynamic incentives for effort, especially when equilibrium effort is not expected to be too large. To see why, recall that equilibria in baseline and type-extension models can be described by triples (E^s, C^s, R^s_1) and (E^s_D, C^s, R^s_1), respectively. If $(E^s, C^s, R^s_1) = (E^s_D, C^s, R^s_1)$, then the difference in posteriors in the baseline model ($\bar{\mu}$ as defined in main text) is always larger than then difference in posteriors in the extension ($\bar{\mu}$ as defined above). So the stakes of reporting game are smaller in the extension and than in the baseline model all else equal. This implies larger dynamic effort incentives in the baseline model. However, as enforcement effort gets larger, criminal activity decreases, so C^s is quite small on the right-hand-side of this panel. This effect is particularly strong in the baseline model, where there are no lazy types who

Figure C.2: Equilibrium effort, criminal activity, and misreporting across the models.



Notes: Red, dashed lines denote the baseline model and blue solid lines denote the type-extension model. For the type-extension model, the left panel graphs the diligent type's equilibrium effort, E_D^s. Example generated using the same assumptions as Figure C.1.

always exert low effort, which is why the baseline model has smaller dynamic incentives for effort with large E^s.

Second, in the type extension, the dynamic incentives for effort are largest when the prior is neither too large nor too small. When the prior is extreme, reporting no crime does not substantively change third-party beliefs about the agent's type. Furthermore, comparing the extreme cases, there are larger dynamic incentives when diligent types are rare (i.e., $\pi = 0.2$) than when diligent types dominate (i.e., $\pi = 0.8$). This asymmetry emerges via the target's best response. All else equal, the rate of criminal activity C^s is larger when π is close to zero rather than close to one, which implies the dynamic incentives for effort will be larger in the former case.

As a consequence of the above discussion, equilibrium effort can be larger or smaller in the type-extension than in the baseline model, and this will depend on the prior. To see this, Figure C.2 graphs the equilibrium probabilities of effort, criminal activity, and misreporting. Note that, in the left panel, we graph the probability of effort of a diligent type, E_D^s in blue. This probability is larger than the probability of effort in the baseline model when π is moderate, where dynamic incentives for effort are the largest. In contrast, this probability is smaller than the probability of effort in the baseline model when π is extreme (i.e., close to 0 or 1). For these values, dynamic incentives for effort are the smallest.

D Analysis with probabilistic crime production

In this Appendix, we assume that crime is produced probabistically. One can motivate this as follows: with probability α, $x = (1-e)c$; with complementary probability $1-\alpha$, $x = c$. Thus, we interpret α as how sensitive the crime outcome is to enforcement effort. The baseline model covered the case where $\alpha = 1$, when high effort surely prevented crime. As α becomes smaller, high enforcement effort does not necessarily prevent crime from occurring, and illicit activity becomes more important to crime production. We can write the probability of a crime outcome as

$$\Pr(x=1 \mid e,c) = \begin{cases} c & \text{if } e=0 \\ (1-\alpha)c & \text{if } e=1 \end{cases}.$$

We will characterize "full support" equilibria, i.e.: (i) $E^s \in (0,1)$ and $C^s > 0$, (ii) after $x = 0$, the agent reports truthfully by sending $\tilde{x} = 0$, and (iii) after $x = 1$, the agent potentially reclassifies the crime statistic by sending $\tilde{x} = 0$.

To begin the analysis, note that after observing $\tilde{x} = 1$, we know the true outcome was $x = 1$. We need to compute:

$$\begin{aligned}
\Pr(e=1 \mid \tilde{x}=1) &= \frac{\Pr(\tilde{x}=1 \mid e=1)\Pr(e=1)}{\Pr(\tilde{x}=1)} \\
&= \frac{\Pr(\tilde{x}=1 \mid e=1)E^s}{\Pr(\tilde{x}=1 \mid x=0)\Pr(x=0)+\Pr(\tilde{x}=1 \mid x=1)\Pr(x=1)} \\
&= \frac{[\Pr(\tilde{x}=1|e=1,x=0)\Pr(x=0|e=1)+\Pr(\tilde{x}=1|e=1,x=1)\Pr(x=1|e=1)]E^s}{\Pr(\tilde{x}=1 \mid x=1)\Pr(x=1)} \\
&= \frac{(1-R_1^s)C^s(1-\alpha)E^s}{(1-R_1^s)C^s[(1-\alpha)E^s+(1-E^s)]} \equiv b_1.
\end{aligned}$$

Above, we need to invoke $\Pr(x=1) = C^s[(1-\alpha)E^s + (1-E^s)]$. Notice that $\alpha = 1$ implies $\Pr(e = 1 \mid \tilde{x} = 1) = 0$. In addition, $\alpha = 0$ implies $\Pr(e = 1 \mid \tilde{x} = 1) = E^s$. So α is moderating the informativeness of the crime report.

After observing $\tilde{x} = 0$, we need to compute:

$$\begin{aligned}
\Pr(e = 1 \mid \tilde{x} = 0) &= \frac{\Pr(\tilde{x} = 0 \mid e = 1) \Pr(e = 1)}{\Pr(\tilde{x} = 0)} \\
&= \frac{\Pr(\tilde{x} = 0 \mid e = 1) E^s}{\Pr(\tilde{x} = 0 \mid x = 0) \Pr(x = 0) + \Pr(\tilde{x} = 0 \mid x = 1) \Pr(x = 1)} \\
&= \frac{[\Pr(\tilde{x} = 0 | e = 1, x = 0) \Pr(x = 0 | e = 1) + \Pr(\tilde{x} = 0 | e = 1, x = 1) \Pr(x = 1 | e = 1)] E^s}{\Pr(x = 0) + R_1^s \Pr(x = 1)} \\
&= \frac{[\Pr(x = 0 | e = 1) + R_1^s \Pr(x = 1 | e = 1)] E^s}{\Pr(x = 0) + R_1^s \Pr(x = 1)} \\
&= \frac{[1 - C^s(1 - \alpha)(1 - R_1^s)] E^s}{1 - C^s(1 - R_1^s)(1 - E^s \alpha)} \equiv b_0
\end{aligned}$$

Notice that b_0 is strictly increasing in α, so the no-crime report is a more convincing signal of effort when crime is more sensitive to effort.

For the equilibrium threshold of lying, we need to compute the difference in posteriors:

$$\begin{aligned}
\bar{\mu}(E^s, C^s, R_1^s) &\equiv b_0 - b_1 \\
&= \frac{E^s(1 - E^s)\alpha}{(1 - E^s\alpha)(1 - C^s(1 - R_1^s)(1 - E^s\alpha))}
\end{aligned}$$

Notice $\bar{\mu} > 0$; $\bar{\mu}$ is strictly increasing in C^s and strictly decreasing in R_1^s. For E^s close to zero, $\bar{\mu}$ is strictly increasing in E^s. For E^s close to one, $\bar{\mu}$ is strictly decreasing. Finally, $\bar{\mu}$ is increasing in the crime-outcome sensitivity to effort, α.

The agent manipulates if and essentially only if $\eta < \hat{\eta}^*$, and the cutpoint $\hat{\eta}^*$ solves

$$\bar{\mu}(E^s, C^s, H(\hat{\eta}^*)) = \hat{\eta}^*$$

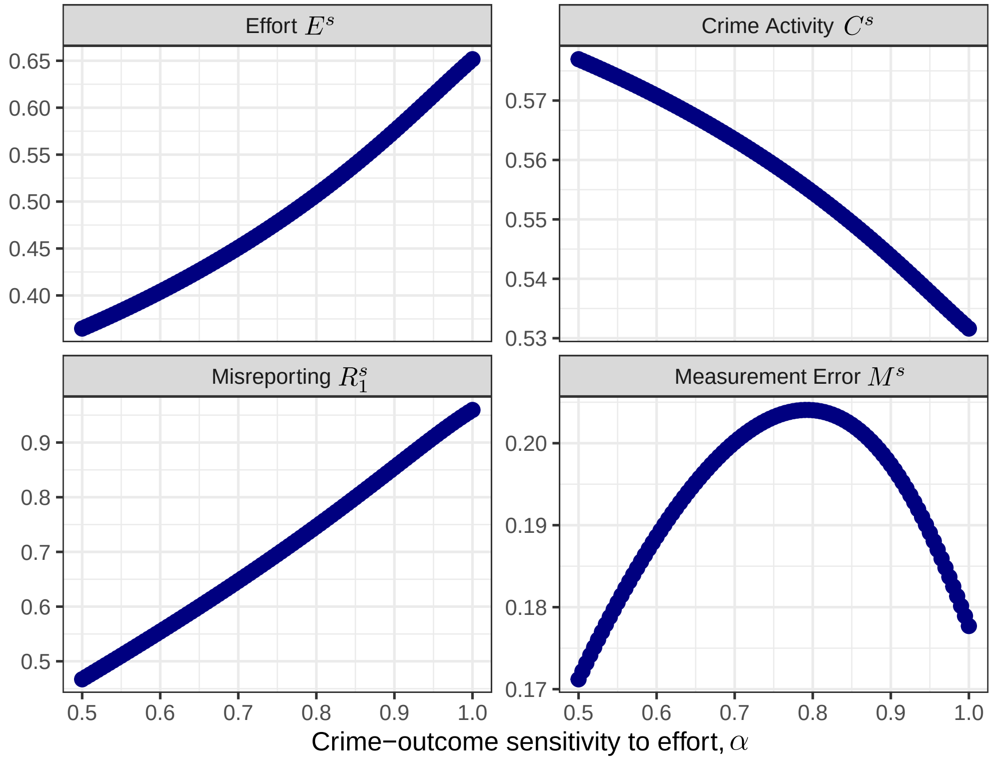
Thus $R_1^s = H(\hat{\eta}^*)$.

For the enforcement game, the Target has the same payoffs so $C^s = G(1 - E^s)$. To complete the construction of the equilibrium, focus on the Agent's payoff. When A exerts effort, she expects to receive $\beta C^s - \rho$ in the enforcement stage. Moreover, with probability $1 - C^s$ there is no illicit activity leading to outcome $x = 0$. With probability $C^s \alpha$, there is illicit activity but no crime, leading to $x = 0$. With probability $C^s(1 - \alpha)$, there is illicit activity and the crime outcome is $x = 1$. For each case in which $x = 0$, reporting payoffs are b_0. For each case in which $x = 1$, reporting payoffs are $(1 - R_1^s)b_1 + R_1^s(b_0 - \mathbb{E}[\eta | \eta < \bar{\mu}])$. Hence, after exerting effort, A's expected payoffs are

$$\beta C^s - \rho + (1 - C^s + C^s\alpha)b_0 + C^s(1 - \alpha)[(1 - R_1^s)b_1 + R_1^s(b_0 - \mathbb{E}[\eta | \eta < \bar{\mu}])].$$

When not exerting effort, A's payoff in the enforcement game is 0. There is no crime if

Figure D.1: Equilibrium quantities as a function of crime-outcome sensitivity to effort



Notes: Example generated assuming $\beta = 0.4$, $\rho \sim \mathcal{N}(0.3, 0.1)$, $\gamma \sim \mathcal{N}(0.15, 2.5)$, and $\eta \sim \mathcal{B}(1, 3)$. Recall that in this extension $\Pr(x = 1 \mid e, c) = e(1-\alpha)c + (1-e)c$, so $(1-\alpha)c$ is the probability that crime occurs after high enforcement effort. The baseline model assumed $\alpha = 1$.

$C^s = 0$. If $C^s = 1$, then crime occurs. Thus, after low effort, A's expected payoffs are

$$(1 - C^s)b_0 + C^s((1 - R_1^s)b_1 + R_1^s(b_0 - \mathbb{E}[\eta|\eta < \bar{\mu}])).$$

Comparing the preceding expressions, A exerts effort if and only if

$$\rho < C^s \left[\beta + \alpha\Psi(E^s, C^s)\right]$$

where

$$\Psi(E^s, C^s) = \bar{\mu}(1 - R_1^s) + R_1^s\mathbb{E}[\eta|\eta < \bar{\mu}]$$

This is analogous to the case analyzed in the main text. Note that crime-outcome sensitivity α affects both the general incentives of the reporting stage (it multiplies Ψ) and the equilibrium misreporting threshold (via $\bar{\mu}$, which is increasing in α).

To conclude this section, Figure D.1 presents a numerical example to illustrate comparative statics with respect to how sensitive crime is to agent effort, i.e., α. Notice that as α

moves away from one and closer to zero, effort decreases. As discussed above, this is driven by two forces: smaller reputational benefits in $\bar{\mu}$ and smaller direct effects on the crime outcome. As effort decreases crime increases. Both of these forces increase measurement error because they increase the probability of the crime outcome occurring. Finally, as α becomes smaller, misreporting decreases. With the probability of agent effort decreasing and the no-crime report becoming less indicative off high effort, reputational benefits decrease, leading to less misreporting. This effect reduces measurement error. Overall, these two countervailing forces lead to the non-monotonic relationship between α and measurement error in the figure.

E Full-support equilibria with remedial policing

Setup

The setup of the game is the same as above, except for the definition of the crime statistic x. Here, we assume that $x = ec$.

Full-support Equilibrium

Definition E.1. *An equilibrium (s, b) has full support if $E^s \in (0, 1)$, $C^s > 0$, $R_1^s = 0$ and $R_0^s \in (0, 1)$.*

The only difference concerns the misreporting probabilities: while the main text features $R_0^s = 0$ and $R_1^s \in (0, 1)$, here we consider data manipulation after no crime occurred.

As in the preventative policing model, our focus on full-support equilibria can to some extent by justified by observing that both crime and no-crime reports should be sent in equilibrium, see Lemma A.1.

Equilibrium Characterization

In a full-support equilibrium, posterior beliefs can be derived as follows. After $\tilde{x} = 1$, we have:

$$\begin{aligned}\Pr(e = 1|\tilde{x} = 1) = b_1 &= \frac{\Pr(e = 1)\Pr(\tilde{x} = 1|e = 1)}{\Pr(\tilde{x} = 1)}\\ &= \frac{E^s\left[C^s + (1 - C^s)R_0^s\right]}{E^sC^s + (1 - E^sC^s)R_0^s}.\end{aligned}$$

Note that b_1 is increasing in E^s, increasing in C, and decreasing in R_0.

After $\tilde{x} = 0$, we have:

$$\begin{aligned}\Pr(e = 1|\tilde{x} = 0) = b_0 &= \frac{\Pr(e = 1)\Pr(\tilde{x} = 1|e = 1)}{\Pr(\tilde{x} = 0)} \\ &= \frac{E^s\left[C^s 0 + (1 - C^s)(1 - R_0^s)\right]}{C^s E^s 0 + (1 - C^s E^s)(1 - R_0^s)} \\ &= \frac{E^s(1 - C^s)}{1 - E^s C^s}.\end{aligned}$$

Intuitively, when $\tilde{x} = 0$ is reported, the principal can infer that $x = 0$ indeed occurred, which is imperfectly informative about the agent's actual effort choice e. By inspection, b_0 is increasing in E^s, decreasing in C^s, and does not depend on R_0^s.

Note that $b_1 > b_0$ in a full-support equilibrium (being defined as having $E^s < 1$). For equilibrium behavior, a key quantity will be the difference in these posterior beliefs, i.e.,

$$\bar{\mu}(E^s, C^s, R_0^s) \equiv b_1 - b_0.$$

Note the following:

Lemma E.1. *$\bar{\mu}$ is increasing in C^s and decreasing in R_0^s. Moreover, $\bar{\mu}$ is concave in E^s and satisfies:*

$$\left.\frac{\partial \bar{\mu}}{\partial E^s}\right|_{E^s=0} > 0 \quad \textit{and} \quad \left.\frac{\partial \bar{\mu}}{\partial E^s}\right|_{E^s=1} < 0.$$

Proof. Differentiating $\bar{\mu}$ with respect to E^s yields:

$$\frac{\partial \bar{\mu}}{\partial E^s} = \frac{R_0^s\left[C^s + (1 - C^s)R_0^s\right]}{\Pr(\tilde{x} = 1)^2} - \frac{1 - C^s}{\Pr(\tilde{x} = 0)^2}$$

Evaluting this at $E^s = 0$ yields $\frac{C^s}{R_0} > 0$ while evaluating this at $E^s = 1$ yields $\frac{-C^s\left(1-R_0^s\right)}{C^s+(1-C^s)R_0^s} < 0$.

Moreover, we have:

$$\frac{\partial^2 \bar{\mu}}{\partial (E^s)^2} = \frac{-2R_0^s\left[C^s + (1 - C^s)R_0\right]}{\Pr(\tilde{x} = 1)^3}\frac{\partial \Pr(\tilde{x} = 1)}{\partial E^s} - \left[\frac{-2(1 - C^s)}{\Pr(\tilde{x} = 0)^3}\frac{\partial \Pr(\tilde{x} = 0)}{\partial E^s}\right]$$

Since $\frac{\partial \Pr(\tilde{x}=1)}{\partial E^s} = C^s(1 - R_0^s) > 0$ and $\frac{\partial \Pr(\tilde{x}=0)}{\partial E^s} = -C^s < 0$, the result follows. □

At the reporting stage, when $x = 0$, R_0^s is determined as follows: the expected utility of choosing $\tilde{x} = 0$ is b_0 while the expected utility of $\tilde{x} = 1$ is $b_1 - \eta$. A threshold strategy is optimal. The equilibrium threshold $\hat{\eta}^*$ is given by the solution to

$$\overline{\mu}(E^s, C^s, H(\hat{\eta}^*)) = \hat{\eta}^* \tag{E.1}$$

The left-hand side of the preceding equality is decreasing in $\hat{\eta}$ while the right-hand side is increasing in $\hat{\eta}$, so the solution is unique. Note that by the IFT, $\hat{\eta}^*$ is non-monotone in E^s and increasing in C^s. Specifically, we have:

Lemma E.2. *We have:*

$$\frac{\partial \hat{\eta}^*}{\partial E} = -\frac{\frac{\partial \bar{\mu}}{\partial E}}{\frac{\partial \bar{\mu}}{\partial R_0} h(\hat{\eta}) - 1} \quad \textit{and} \quad \frac{\partial \hat{\eta}^*}{\partial C} = -\frac{\frac{\partial \bar{\mu}}{\partial C}}{\frac{\partial \bar{\mu}}{\partial R_0} h(\hat{\eta}) - 1} > 0$$

Proof. Follows directly from the IFT and Lemma E.1. □

At the encounter stage, the best response function of the Target is the same as before:

$$\hat{\gamma}^* = 1 - E^s \tag{E.2}$$

For the agent, the expected utility of effort is:

$$C^s\beta - \rho + \Pr(x = 1|e = 1)b_1 + \Pr(x = 0|e = 1)\left[R_0^s\left(b_1 - \mathbb{E}[\eta|\eta < \hat{\eta}^*]\right) + (1 - R_0^s)b_0\right]$$

Furthermore, the expected utility of no effort is:

$$R_0^s\left(b_1 - \mathbb{E}[\eta|\eta < \hat{\eta}^*]\right) + (1 - R_0^s)b_0$$

Thus, the agent exerts effort if and only if:

$$C^s\left[\beta + \Psi(E^s, C^s)\right] \geq \rho,$$

where

$$\begin{aligned}\Psi(E^s, C^s) &\equiv (1 - R_0^s)\bar{\mu}\left(E^s, C^s, H(\hat{\eta}^*)\right) + R_0^s\mathbb{E}[\eta|\eta < \hat{\eta}^*] \\ &= \left(1 - H(\hat{\eta}^*)\right)\hat{\eta}^*) + \int_{\underline{\eta}}^{\hat{\eta}^*} \eta h(\eta) d\eta\end{aligned}$$

Plugging this into the preceding expressions, an equilibrium (s, b) is characterized by a threshold strategy, $\hat{\rho}^*$, that solves

$$\underbrace{G(1 - F(\hat{\rho}))[\beta + \Psi(F(\hat{\rho}), G(1 - F(\hat{\rho})))]}_{\equiv \Lambda(\hat{\rho})} = \hat{\rho}. \tag{E.3}$$

Proposition E.1 summarizes the analysis thus far.

Proposition E.1. *If (s, b) is a full-support equilibrium, then the following hold:*

1. *The agent exerts effort if and only if $\rho < \hat{\rho}^*$ where ρ^* solves Equation E.3, so $E^s = F(\hat{\rho}^*)$.*

2. *The target engages in illicit behavior if and only if $\gamma < \hat{\gamma}^*$ where $\hat{\gamma}^*$ solves Equation E.2, so $C^s = G(\hat{\gamma}^*)$.*

3. *In the reporting subgame, the agent never misclassifies the no-crime statistic $x = 1$, but misclassifies after crime statistic $x = 0$ if and only if $\eta < \hat{\eta}^*$, where $\hat{\eta}^*$ solves Equation E.1. So $R_0^s = H(\hat{\eta}^*)$.*

F Equilibrium uniqueness

F.1 The function Λ is decreasing on a relevant range

Before proceeding, we want to know how Ψ changes as a function of E^s and C^s. Recall that $\Psi(E^s, C^s) = (1 - H(\hat{\eta}^*))\hat{\eta}^* + \int_{\underline{\eta}}^{\hat{\eta}^*} \eta h(\eta) d\eta$. By Leibniz's rule, we can write

$$\begin{aligned}
\frac{\partial \Psi}{\partial C^s} &= -h(\hat{\eta}^*)\frac{\partial \hat{\eta}^*}{\partial C}\hat{\eta}^* + (1 - H(\hat{\eta}^*))\frac{\partial \hat{\eta}^*}{\partial C^s} + \hat{\eta}^* h(\hat{\eta}^*)\frac{\partial \hat{\eta}^*}{\partial C} \\
&= h(\hat{\eta}^*)\frac{\partial \hat{\eta}^*}{\partial C}\left(\hat{\eta}^* - \hat{\eta}^*\right) + (1 - H(\hat{\eta}^*))\frac{\partial \hat{\eta}^*}{\partial C^s} \\
&= (1 - H(\hat{\eta}^*))\frac{\partial \hat{\eta}^*}{\partial C^s} \geq 0.
\end{aligned}$$

Second, a similar analysis shows

$$\frac{\partial \Psi}{\partial E^s} = (1 - H(\hat{\eta}^*))\frac{\partial \hat{\eta}^*}{\partial E^s} \geq 0.$$

The next proposition states a sufficient condition for a unique equilibrium.

Proposition F.1. *There is a unique solution to Equation 3 if*

$$\begin{aligned}
&\min\{g(\gamma) \mid \gamma \in [1 - F(G(1)(\beta + 1)), 1 - F(0)]\} \geq \\
&\quad \max\left\{\frac{1 - G(0)(1 - H(1))}{F(0)(1 - F(0))(1 - H(1))}, \frac{1 - G(0)(1 - H(1))}{F(G(1)(\beta + 1))(1 - F(G(1)(\beta + 1)))(1 - H(1))}\right\}.
\end{aligned}$$

Proof. First, note that right-side of Equation 3 is strictly increasing in ρ. Thus, a sufficient condition for a unique solution is that Λ is decreasing over all feasible equilibrium cutpoints $\hat{\rho}$.

Second, note that derivative of Λ with respect to ρ is

$$\begin{aligned}
\frac{\partial \Lambda}{\partial \hat{\rho}} &= -g(1 - F(\hat{\rho}))f(\hat{\rho})[\beta + \Psi(F(\hat{\rho}), G(1 - F(\hat{\rho})))] + \\
&\qquad G(1 - F(\hat{\rho}))\left[\frac{\partial \Psi}{\partial E^s}f(\hat{\rho}) - \frac{\partial \Psi}{\partial C^s}g(1 - F(\hat{\rho}))f(\hat{\rho})\right].
\end{aligned}$$

To sign this expression, substitute our expressions for $\frac{\partial \Psi}{\partial E^s}$ and $\frac{\partial \Psi}{\partial C^s}$:

$$\frac{\partial \Lambda}{\partial \hat{\rho}} = -g(1-F(\hat{\rho}))f(\hat{\rho})[\beta + \Psi(F(\hat{\rho}), G(1-F(\hat{\rho})))] + \\ G(1-F(\hat{\rho}))f(\hat{\rho})(1-H(\hat{\eta}^*))\left[\frac{\partial \hat{\eta}^*}{\partial E^s} - \frac{\partial \hat{\eta}^*}{\partial C^s} g(1-F(\hat{\rho}))\right].$$

Using Lemma 2, we substitute our expressions for $\frac{\partial \hat{\eta}^*}{\partial E^s}$ and $\frac{\partial \hat{\eta}^*}{\partial C^s}$ to get

$$\frac{\partial \Lambda}{\partial \hat{\rho}} = -g(1-F(\hat{\rho}))f(\hat{\rho})[\beta + \Psi(F(\hat{\rho}), G(1-F(\hat{\rho})))] + \\ G(1-F(\hat{\rho}))f(\hat{\rho})\frac{(1-H(\hat{\eta}^*))}{1-\frac{\partial \bar{\mu}}{\partial R_1^s} h(\hat{\eta}^*)}\left[\frac{\partial \bar{\mu}}{\partial E^s} - \frac{\partial \bar{\mu}}{\partial C^s} g(1-F(\hat{\rho}))\right].$$

Notice $\frac{\partial \bar{\mu}}{\partial E^s} > 0$ and $\frac{\partial \bar{\mu}}{\partial C^s} > 0$. Thus, $\left[\frac{\partial \bar{\mu}}{\partial E^s} - \frac{\partial \bar{\mu}}{\partial C^s} g(1-F(\hat{\rho}))\right] \leq 0$ implies $\frac{\partial \Lambda}{\partial \hat{\rho}} < 0$. Substituting expressions for $\frac{\partial \bar{\mu}}{\partial E^s}$ and $\frac{\partial \bar{\mu}}{\partial C^s}$, $\frac{\partial \Lambda}{\partial \hat{\rho}} < 0$ if

$$g(1-F(\hat{\rho})) \geq \frac{1-C^s(1-R_1^s)}{E^s(1-E^s)(1-R_1^s)} \equiv \delta^g.$$

Careful inspection reveals that δ^g is strictly decreasing in C^s and strictly increasing in R_1^s. Recall $C^s \geq G(0)$ and $R_1^s \leq H(1)$ in any equilibrium (s, b) Thus, a lower bound on δ^g is $\delta^g|_{R_1^s=H(1), C^s=G(0)}$. Likewise, $E^s \in [F(0), F(G(1)(\beta+1))]$ in any equilibrium (s, b), and δ^g is convex in E^s. Thus, a upper bound on δ^g is

$$\max\{\delta^g|_{R_1^s=H(1), C^s=G(0), E^s=F(0)}, \delta^g|_{R_1^s=H(1), C^s=G(0), E^s=F(G(1)(\beta+1))}\},$$

which is the right-hand of the inequality in the statement of the proposition. Finally, when the agent uses threshold strategy $\hat{\rho}$, $E^s = F(\hat{\rho})$ and $E^s \in E^s \in [F(0), F(G(1)(\beta+1))]$ in any equilibrium. So a lower bound for $g(1-F(\hat{\rho}))$ is $\min\{g(\gamma) \mid \gamma \in [1-F(G(1)(\beta+1)), 1-F(0)]\}$, which is the left-and side of the inequality in the statement of the proposition. □

F.2 Uniqueness with enough noise in F, G, and H

Notice that C^s, R_1^s, and E^s serve as the induced probabilities of effort, crime and misclassification. Collect these probabilities in the vector $\pi^s = (R_1^s, C^s, E^s)$. Define the function $\mathcal{E} : [0,1]^3 \to \mathbb{R}^3$ as follows:

$$\mathcal{E}(\pi^s) = \begin{bmatrix} E^s - F(C^s[\beta + \Psi(\pi^s)]) \\ C^s - G(1-E^s) \\ H(\bar{\mu}(\pi^s)) - R_1^s \end{bmatrix},$$

where $\Psi : [0,1]^3 \to \mathbb{R}$ can be expressed as

$$\Psi(\pi^s) = (1 - R_1^s)\bar{\mu}(\pi^s) + \int_{\underline{\eta}}^{\bar{\mu}(\pi^s)} \eta h(\eta) d\eta.$$

An equilibrium is π^s such that $\mathcal{E}(\pi^s) = 0$. Given a solution π^s such that $\mathcal{E}(\pi^s) = 0$, one can construct the thresholds in Proposition 1 because $E^s = F(\hat{\rho}^*)$, $C^s = G(\hat{\gamma}^*)$, and $R^s = H(\hat{\eta}^*)$. Focusing on Ψ, we can write the partial derivatives of Ψ as follows:

$$\begin{aligned}
\frac{\partial \Psi}{\partial E^s} &= [(1 - R_1^s) + \bar{\mu}(\pi^s)h(\bar{\mu}(\pi^s))]\frac{\partial \bar{\mu}}{\partial E^s} \geq 0 \\
\frac{\partial \Psi}{\partial C^s} &= [(1 - R_1^s) + \bar{\mu}(\pi^s)h(\bar{\mu}(\pi^s))]\frac{\partial \bar{\mu}}{\partial C^s} \geq 0 \\
\frac{\partial \Psi}{\partial R_1^s} &= [(1 - R_1^s) + \bar{\mu}(\pi^s)h(\bar{\mu}(\pi^s))]\frac{\partial \bar{\mu}}{\partial R_1^s} - \bar{\mu}(\pi^s) \leq 0.
\end{aligned}$$

Notice the inequalities above hold strictly when either $R_s^1 < 1$, or $E^s > 0$ with $h(\mu) > 0$ for all $\mu \in [0,1]$. With these partial derivatives in hand, we can write the Jacobian of $\mathcal{E}$ as

$$J\mathcal{E} = \begin{bmatrix} -f(C^s(\beta+\Psi))C^s\frac{\partial \Psi}{\partial R_1^s} & -f(C^s(\beta+\Psi))\left[(\beta+\Psi) + C^s\frac{\Psi}{\partial C^s}\right] & 1 - f(C^s(\beta+\Psi))C^s\frac{\partial \Psi}{\partial E^s} \\ 0 & 1 & g(1-E^s) \\ -1 + h(\bar{\mu})\frac{\partial \bar{\mu}}{\partial R_1^s} & h(\bar{\mu})\frac{\partial \bar{\mu}}{\partial C^s} & h(\bar{\mu})\frac{\partial \bar{\mu}}{\partial E^s} \end{bmatrix}.$$

Notice the Jacobian $J\mathcal{E}$ is a 3×3 matrix. For $N \subset \{1,2,3\}$, let $J\mathcal{E}_{-N}$ denote the principal submatrix of $J\mathcal{E}$ formed by deleting the rows and columns in N. So the diagonal of $J\mathcal{E}$ comprises the values $J\mathcal{E}_{-\{2,3\}}$, $J\mathcal{E}_{-\{1,3\}}$, and $J\mathcal{E}_{-\{1,2\}}$. Recall that a matrix is a P-matrix if all of its if all its principal minors (determinants of a principal submatrix) are positive.

Lemma F.1. *The Jaobian $J\mathcal{E}$ is a P-matrix at $\pi^s = (R_1^s, C^s, E^s) \in [0,1]^3$ if the following conditions hold.*

1. $h(\mu) > 0$ *for all* $\mu \in [0,1]$
2. $f(C^s(\beta + \Psi)) > 0$
3. $R_1^s < 1$
4. $\frac{\partial \bar{\mu}}{\partial E^s} > g(1 - E^s)\frac{\partial \bar{\mu}}{\partial C^s}$
5. $1 - f(C^s(\beta + \Psi))C^s\frac{\partial \Psi}{\partial E^s} \geq 0$

Proof. Fix $\pi^s \in D$. We need to show that the principal minors of $J\mathcal{E}$ are positive. Under the five conditions above, the sign of $J\mathcal{E}$ can be written as

$$\operatorname{sgn}(J\mathcal{E}) = \begin{bmatrix} 1 & -1 & \{0,1\} \\ 0 & 1 & \{0,1\} \\ -1 & 1 & 1 \end{bmatrix}.$$

Above, we use $\{0,1\}$ to represent the possibility that an entry is either 0 or positive. With this in hand, we can consider the 7 principal submatrices of $J\mathcal{E}$.

- Starting with the diagonals, $J\mathcal{E}_{-\{2,3\}}$, $J\mathcal{E}_{-\{1,3\}}$, and $J\mathcal{E}_{-\{1,2\}}$. Notice $J\mathcal{E}_{-\{1,3\}} = 1 > 0$. $J\mathcal{E}_{-\{1,2\}} = h(\bar{\mu})\frac{\partial\bar{\mu}}{\partial E^s}$, which is greater than 0 if $h(\mu) > 0$ for all $\mu \in [0,1]$, as covered in condition 1. Finally, $J\mathcal{E}_{-\{2,3\}} = -f(C^s(\beta+\Psi))C^s\frac{\partial\Psi}{\partial R_1^s}$. This value is strictly positive by conditions 2 and 3.

- Now focus on $J\mathcal{E}_{-\{1\}}$. The determinant of this matrix is

$$h(\bar{\mu})\frac{\partial\bar{\mu}}{\partial E^s} - g(1-E^s)h(\bar{\mu})\frac{\partial\bar{\mu}}{\partial C^s}.$$

  Because $h(\mu) > 0$ for all $\mu \in [0,1]$, this expression is strictly positive when $\frac{\partial\bar{\mu}}{\partial E^s} > g(1-E^s)\frac{\partial\bar{\mu}}{\partial C^s}$, which is covered in condition 4.

- Now focus on $J\mathcal{E}_{-\{2\}}$. Examining $\operatorname{sgn}(J\mathcal{E})$ reveals that the determinant of $J\mathcal{E}_{-\{2\}}$ is strictly positive because $1 - f(C^s(\beta+\Psi))C^s\frac{\partial\Psi}{\partial E^s} \geq 0$ from condition 5.

- Now focus on $J\mathcal{E}_{-\{3\}}$. The sign matrix reveals that this determinant is strictly positive.

- Finally, we need to compute the determinant of $J\mathcal{E}$. We can do this as follows:

$$\overbrace{J\mathcal{E}_{-\{2,3\}} \cdot \left|J\mathcal{E}_{-\{1\}}\right|}^{>0} - \overbrace{\left(-f(C^s(\beta+\Psi))\left[(\beta+\Psi) + C^s\frac{\Psi}{\partial C^s}\right]\right)\left|\begin{bmatrix} 0 & g(1-E^s) \\ -1 + h(\bar{\mu})\frac{\partial\bar{\mu}}{\partial R_1^s} & h(\bar{\mu})\frac{\partial\bar{\mu}}{\partial E^s} \end{bmatrix}\right|}^{\leq 0} +$$
$$\underbrace{\left(1 - f(C^s(\beta+\Psi))C^s\frac{\partial\Psi}{\partial E^s}\right)}_{\geq \text{ by condition 5}}\underbrace{\left|\begin{bmatrix} 0 & 1 \\ -1 + h(\bar{\mu})\frac{\partial\bar{\mu}}{\partial R_1^s} & h(\bar{\mu})\frac{\partial\bar{\mu}}{\partial C^s} \end{bmatrix}\right|}_{>0}.$$

Thus, the determinant of $J\mathcal{E}$ is positive given the five conditions above. □

Besides condition 1, the conditions in Lemma F.1 are restrictions on endogenous quantifies of interest. To rectify this, recall $R_1^s \leq H(1)$ at any solution π^s (equilibrium) such that $\mathcal{E}(\pi^s) = 0$. Likewise, $C^s = G(1-E^s)$ and $E^s \in [0,1]$, so $C^s \in [G(0), G(1)]$ at any solution π^s (equilibrium) such that $\mathcal{E}(\pi^s) = 0$. Finally, $E^s = F(C^s(\beta+\Psi))$, where $0 \leq C^s(\beta+\Psi) \leq \beta+1$. Hence, $E^s \in [F(0), F(G(1)(\beta+1))]$ at any solution π^s

(equilibrium) such that $\mathcal{E}(\pi^s) = 0$. Thus, we can restrict the domain in Lemma F.1 to $\tilde{D} = [0, H(1)] \times [0, G(1)] \times [F(0), F(G(1)(\beta + 1))]$, where $\tilde{D} \subseteq [0,1]^3$. That is, any π^s such that $\mathcal{E}(\pi^s) = 0$ will fall within $\tilde{D}$.

Second, focus on Condition 4. Using our expressions for $\frac{\partial \bar{\mu}}{\partial E^s}$ and $\frac{\partial \bar{\mu}}{\partial C^s}$, we can rewrite this condition as

$$g(1 - E^s) < \frac{1 - C^s(1 - R_1^s)}{E^s(1 - E^s)(1 - R_1^s)} \equiv \delta^g.$$

Recall δ^g is strictly decreasing in C^s and strictly increasing in R_1^s. By construction of the domain $\tilde{D}$, $R_1^s \geq 0$ and $C^s \leq G(1)$. Likewise, it is convex in E^s, with a global minimum at $E^s = \frac{1}{2}$ for all C^s, R_1^s. As such, δ^g is bounded below by $\delta^g \geq 4(1 - G(1))$.

Third, focus on Condition 5, which is equivalent to

$$f(C^s(\beta + \psi)) \leq \left(C^s[(1 - R_1^s) + \bar{\mu}(\pi^s) h(\bar{\mu}(\pi^s))] \frac{\partial \bar{\mu}}{\partial E^s} \right)^{-1} \tag{F.1}$$

We want to find a lower bound for the right-hand side of Equation F.1. Here, note that

$$\begin{aligned} \left(C^s[(1 - R_1^s) + \bar{\mu}(\pi^s) h(\bar{\mu}(\pi^s))] \frac{\partial \bar{\mu}}{\partial E^s} \right)^{-1} &\geq \left(G(1)[1 + h(\bar{\mu}(\pi^s))] \frac{\partial \bar{\mu}}{\partial E^s} \right)^{-1} \\ &\geq \left(\frac{1}{G(1)} \right) \underbrace{\frac{(1 - C^s(1 - E^s)(1 - R_1^s))^2}{1 - C^s(1 - R_1^s)}}_{\equiv \delta^f}. \end{aligned}$$

Notice δ^f is a decreasing function of E^s. When $E^s \geq \frac{1}{2}$, δ^f, as a function of $C^s(1 - R_1^s)$, is minimized at $C^s(1 - R_1^s) = 0$, in which case the minimum is 1. Furthermore, when $E^s < \frac{1}{2}$, δ^f is minimized, as a function of $C^s(1 - R_1^s)$, at $C^s(1 - R_1^s) = \frac{1 - 2E^s}{1 - E^s}$, in which case the minimum is $4(1 - E^s)E^s$. This quantity is less than 1 and is strictly increasing in E^s. Thus, assuming $F(0) < \frac{1}{2}$, a lower bound on the right-hand side of Equation F.1 is $\frac{4(1 - F(0))F(0)}{G(1)}$. Putting these three results together, we can state a new result—along the lines of Lemma F.1—that just has restrictions on the primitives and guarantees the Jacobian $J\mathcal{E}$ is a P-matrix for all potential solutions $\pi^s \in \tilde{D}$.

Lemma F.2. *Define $\tilde{D} = [0, H(1)] \times [0, G(1)] \times [F(0), F(G(1)(\beta + 1))]$. The Jacobian $J\mathcal{E}$ is a P-matrix for all $\pi^s = (R_1^s, C^s, E^s) \in \tilde{D}$ if the following conditions hold.*

1. *$f(\rho) > 0$ for all $\rho \in [0, G(1)(\beta + 1)]$*
2. *$h(\mu) > 0$ for all $\mu \in [0, 1]$*
3. *$H(1) < 1$*
4. *$g(\gamma) < 4(1 - G(1))$ for all $\gamma \in [1 - F(G(1)(\beta + 1)), 1 - F(0)]$.*
5. *$F(0) \leq \frac{1}{2}$ and $f(\rho) \leq \frac{4(1 - F(0))F(0)}{G(1)}$ for all $\rho \in [0, G(1)(\beta + 1)]$.*

The next proposition shows that the conditions in Lemma F.2 guarantee a unique equilibrium.

Proposition F.2. *There is a unique equilibrium, i.e., a unique solution to $\mathcal{E}(\pi^s) = 0$ if the following conditions hold.*

1. $f(\rho) > 0$ *for all* $\rho \in [0, G(1)(\beta + 1)]$
2. $h(\mu) > 0$ *for all* $\mu \in [0, 1]$
3. $H(1) < 1$
4. $g(\gamma) < 4(1 - G(1))$ *for all* $\gamma \in [1 - F(G(1)(\beta + 1)), 1 - F(0)]$.
5. $F(0) \leq \frac{1}{2}$ *and* $f(\rho) \leq \frac{4(1-F(0))F(0)}{G(1)}$ *for all* $\rho \in [0, G(1)(\beta + 1)]$

Proof. Recall that π^s is a solution to $\mathcal{E}(\pi^s) = 0$ only if $\pi^s \in \tilde{D}$. Because $\mathcal{E}$ is differentiable and we can restrict its domain to a Cartisan product of intervals, i.e., $\tilde{D} = [0, H(1)] \times [0, G(1)] \times [F(0), F(G(1)(\beta+1))]$, Gale and Nikaido's (1965) theorem implies that a solution in $\tilde{D}$ must be unique because $J\mathcal{E}$ is a P-matrix at all values $\pi^s \in \tilde{D}$ by Lemma F.2. □

The five conditions in Proposition F.2 can be interpreted as requiring enough noise in the random variables ρ, γ, and η. Conditions 1 and 2 require positive density over a relevant interval. Condition 3 requires that lying costs can be large enough. Condition 4 says the density from which opportunity costs are drawn from needs to be sufficiently flat. Condition 5 says the density from which effort costs are drawn from needs to be sufficiently flat.