Enforcement and Compliance in an Uncertain World:

An Experimental Investigation

Eric S. Dickson
New York University

Sanford C. Gordon
New York University

Gregory A. Huber
Yale University

February 28, 2008

Abstract

Governments are charged with monitoring citizens' compliance with prescribed behavioral standards and punishing noncompliance. Flaws in information available to enforcing agents, however, may lead to subsequent enforcement errors, eroding government authority and undermining incentives for compliance. We explore these concepts in a laboratory experiment. A "monitor" makes punishment decisions after receiving noisy signals about individuals' choices to contribute to a public good. We find that the possibility of wrongly accusatory signals has a more deleterious effect on contribution levels than the possibility of wrongly exculpatory signals, contrary to a rational-choice benchmark model predicting equal effects. We trace this across-treatment difference to a "false positives trap": when members of a largely compliant population are sometimes incorrectly accused, some will be unjustly punished if enforcement power is employed, but non-compliant individuals will escape punishment if that power is abdicated. Either kind of error discourages compliance. An additional treatment demonstrates that the effectiveness of a given enforcement institution may vary, depending on its origins. We consider implications of our findings for theories of deterrence, fairness, and institutional legitimacy.

1 Introduction

One of the core functions of government is to monitor citizens' compliance with prescribed standards of "good" behavior and, in the event of noncompliance, to punish misbehavior. In a complex world, however, information available about citizen behavior is not always accurate. Government personnel, charged with the task of punishing violators, may sometimes encounter "false positive" signals, which erroneously identify compliant citizens as noncompliant. Alternatively, they may observe "false negative" signals, which incorrectly identify noncompliant citizens as compliant. Systematic errors in the *information* available to government agents may be critically important, insofar as they lead to errors in their *enforcement decisions*. For example, false positive signals may lead agents to commit "Type I errors," wrongly punishing the innocent, while false negative signals may sometimes lead to "Type II errors," failing to punish the guilty.[1]

Ultimately, the effectiveness of an enforcement regime in encouraging socially-desirable behavior may depend not only on its efficiency but also on the extent to which the exercise of sanctioning authority is perceived as just or fair. Errors in enforcement may, of course, reduce citizens' perceptions of the regime as being either efficient *or* just. If Type I and Type II errors affect these perceptions differently, governments may face fundamental tradeoffs when adopting approaches that make one or the other more or less likely. Understanding these tradeoffs in turn requires uncovering the precise relationship between the threats these errors pose and the behavior of citizens and enforcers alike.

How does the potential for false positive and false negative signals about citizen behavior affect the dynamics of enforcement and compliance? Is authority eroded by such imperfections in the information available to enforcers? If so, do false positive and false negative signals about citizens have comparable, or different, effects? Models of optimal enforcement from law and economics suggest that both types of flawed information and the errors they might induce can undermine the deterrent value of an enforcement regime – false positives and Type I errors by lowering the expected benefit of compliance, and false negatives and Type II errors by raising the expected benefit of noncompliance (Png 1986; Kaplow and Shavell 1994; Polinsky and Shavell 2000, 60-61). Indeed, in a simple rational-choice framework, these kinds of error are effectively substitutes

[1] Throughout, our use of the decision-theoretic terminology of false positives, false negatives, and Type I and Type II errors implies a null hypothesis of innocence, following the convention of "innocent until proven guilty."

for one another, having symmetric effects on citizens' incentives to comply. This view, however, conflicts with a deeply entrenched jurisprudential norm that punishing the innocent is inherently more costly than acquitting the guilty.[2] This norm, often justified by appeals to fairness rather than to efficiency, is summarized in Blackstone's (1769, 352) well-known dictum that, "it is better that ten guilty persons escape, than that one innocent suffer."

In this paper, we explore the problem of enforcement error, and its implications for the legitimacy of governing regimes, in the context of a laboratory experiment on public goods provision. In our experimental scenario, individuals within a group must decide whether or not to contribute to a public good; a "monitor" external to the group then chooses whether or not to punish each individual after receiving noisy information about each individual's contribution decision. Our experiment varies the nature of the uncertainty in the monitor's information across different treatments. Monitors in our False Positives Treatment receive some false positive signals (but no false negatives), whereas monitors in our False Negatives Treatment receive some false negative signals (but no false positives). This design offers us the opportunity to assess the *separate* effects of false positive and false negative signals about citizen behavior on the dynamics of enforcement and compliance. Contrary to the rational-choice benchmark, we find that substantial differences exist between regimes prone to false positive signals and those prone to false negative signals.

Our study also identifies a fundamental challenge in governance that emerges when enforcers receive information that is known to be contaminated with false positive signals about non-compliance. In our experiment, enforcers who are given such information are often reluctant to punish group members, because of the likelihood of committing Type I errors – wrongly punishing those who had in fact contributed to the public good. This reluctance, however, makes monitors susceptible to committing Type II errors – that is, failing to punish non-contributors. Because both Type I and Type II errors by enforcers are estimated to reduce group members' propensities to contribute, monitors under such conditions are caught in what we refer to as a *false positives trap*. The logic of this trap makes high levels of compliance difficult to sustain. Interestingly, there is no corresponding false negatives trap; in the presence of information contaminated only with false negatives, monitors receive few erroneous signals when compliance levels are generally high, and will never

[2]Gordon and Huber (2002) develop a model in which citizens vary in the intensity of this preference in the context of criminal justice. For a discussion of the tradeoffs between Type I and Type II errors in other policy environments, see Canes-Wrone and Shotts (2007).

err by choosing to punish when they *do* receive an accusatory signal.

The logic of the false positives trap suggests that there is a benefit to adopting governing institutions that minimize the possibility of false positive signals about citizen behavior, even if doing so means that false negative signals become more likely. This intuition suggests that Blackstone's aversion to Type I errors – wrongly punishing the guilty – should be bound up in practice with an aversion to monitoring regimes that make innocent citizens *appear* guilty, for reasons of efficiency as well as fairness. The benefits of adopting institutions that minimize false positive signals about citizens are particularly clear-cut when there are strong reasons to believe that underlying compliance rates are high. Our work also suggests the value of selecting governing agents who optimally use the information that is available, taking into account the baseline rates of compliance rather than relying on potentially erroneous signals alone.

We also implement a Technology Choice Treatment in which monitors can choose the nature of the information they receive. Depending on their choice, monitors can either receive information contaminated with false positive signals, or can instead receive perfectly accurate information in exchange for giving up a per-period endowment. Monitors who opt into the condition with false positive signals operate within the context of the *same* monitoring institution as in our False Positives Treatment described above. This aspect of our design allows us to explore whether behavior may differ within the context of a given institution, depending on the nature of that institution's origins, offering a novel empirical window on institutional legitimacy.[3] We find that groups in which the monitor sacrifices perfect accuracy experience *initially* lower contributions than in the (exogenous) False Positives Treatment, but these differences quickly disappear, in part as a consequence of the greater willingness of monitors in the Technology Choice Treatment to punish.

Our experiment contributes to an extensive existing literature on the provision of public goods. Numerous studies have explored the effects of different institutional and structural factors affecting the level of public goods contributions (see, e.g., Ledyard (1995) for a summary of this literature). More recently, a number of studies have explored the effect of punishment on public goods contributions (e.g., Fehr and Gaechter 2000; Bochet, Page, and Putterman 2006) in a variety of experimental settings. Our research adds to this literature by exploring the provision of public

[3]Dal Bo, Foster, and Putterman (2007) explore the idea of *democratic* legitimacy using similarly motivated experimental methods.

goods when an *external* enforcer makes punishment decisions based on systematically *error-prone* information about individual contribution decisions.

Other scholarship has also examined the relationship between individual-level compliance, enforcement errors, and institutional legitimacy from a variety of theoretical and empirical perspectives. Political scientists (Scholz and Pinney 1995; Scholz and Lubell 1998) and social psychologists (e.g., Tyler 1990) alike have found that individuals are more likely to comply with government authority when they believe that others are more likely to do so, when they believe that the enforcer's intentions are noble, and when the enforcer has a reputation for making seemingly reasonable and fair decisions. In line with these intuitions, Scott and Grasmick (1981) find that perceptions of deterrence are more important in shaping taxpayer compliance for individuals who perceive the tax system as unfair, while Feld and Frey (2007, 107) argue that mistreatment by enforcement authorities who take a signal of non-compliance to imply actual non-compliance may offend compliant taxpayers. Alm, McClelland, and Schulze (1999) find that experimental subjects reduced contributions to a public good following the failure of their groups to adopt stronger enforcement via majority rule; in their view, the failure to adopt stronger enforcement institutions led to lower willingness to contribute because that failure itself upset underlying norms. In his review of the normative theoretical literature on trust in government, Hardin (2006, 65-66) notes that a clear distinction must be drawn between mistrust of authority stemming from an authority figure's motivation and those stemming from his or her competence. Our findings shed light on a number of the issues motivating this scholarship.

2 The Rational Agent Benchmark

We begin by describing a formal framework for exploring enforcement and compliance under uncertainty. Specifically, we define the following stage game. There is a monitor m and N members of group g indexed by $i = 1, \ldots, N$.[4] Group members each begin the game with an endowment y_g, which is common to all i. A group member can either contribute his entire endowment to a linear-additive public good ($C_i = 1$) or keep it for himself ($C_i = 0$).[5] Once group members have

[4]Throughout, we employ female pronouns for the monitor and male pronouns for group members.

[5]Because the model is linear, group members would never strictly prefer interior levels of contribution even if these were feasible.

made their allocation choices, the monitor receives noisy signals of the behavior of each group member. Subsequently, she can choose, for each individual i, either to impose a penalty ($P_i = 1$) or not ($P_i = 0$). The magnitude of the penalty is $p \in \mathbb{R}_{++}$.

The payoff to group member i is derived from the endowment if kept, the public good, and the penalty if imposed:

$$u_i(y_g, r, p) = (1 - C_i)y_g + r\left(C_i y_g + \sum_{j \neq i} C_j y_g\right) - P_i p,$$

where $r \in (\frac{1}{N}, 1)$ is the marginal rate of return from contribution.

The payoff to the monitor stems from an endowment y_m and a return on the public good:

$$u_m(y_m, r) = y_m + r\sum_{i=1}^{N} C_i y_g.$$

It is not costly for the monitor to penalize group members, and she is not a residual claimant on penalties. This ensures that she does not have incentives to dissipate rents from office or farm group members for penalties.[6]

For each individual i, the monitor observes a noisy signal $s_i \in \{c, k\}$ of i's behavior. Let $q_1 \in [0, 0.5)$ be the probability the monitor observes an accusatory signal of k ("kept") when i in fact allocated (a false positive signal), and $q_2 \in [0, 0.5)$ the probability the monitor observes an exculpatory signal of c ("contributed") when i in fact kept his endowment (a false negative signal). The information structure is summarized in the following table, the entries of which indicate $\Pr(signal|action)$.

		i's action: Contribute	i's action: Keep
m's signal s_i	c	$1 - q_1$	q_2
	k	q_1	$1 - q_2$

We analyze subgame-perfect equilibria of the stage game. First, consider all subgames following provision decisions by the group members. Because the monitor pays no cost for enforcement, in

[6] In some real-world settings, such conflicts of interest between monitors and group members may of course be critically important, but we do not consider them here.

each of these subgames she will be indifferent between punishing and not punishing each group member. As such, all punishment strategies are sequentially rational.

Next, consider group members' contribution decisions. Suppose that the monitor punishes group member i with probability $\pi_i(s_i)$ upon observing signal s_i, and that each member $j \neq i$ contributes with probability θ_j. Then the expected utilities to member i of contributing and not contributing are given, respectively, by

$$\begin{aligned} E[u_i(C_i = 1)] &= r\left(y_g + \sum_{j \neq i} \theta_j y_g\right) - (1 - q_1)\pi_i(c)p - q_1\pi_i(k)p \\ E[u_i(C_i = 0)] &= y_g + r\sum_{j \neq i} \theta_j y_g - q_2\pi_i(c)p - (1 - q_2)\pi_i(k)p. \end{aligned}$$

Comparing these values, contributing is a best response for group member i if and only if

$$\pi_i(k) - \pi_i(c) \geq \frac{(1 - r)y_g}{(1 - q_1 - q_2)p}. \tag{1}$$

Note that this condition is independent of other group members' decisions ($\theta_{j \neq i}$), and that the right hand side is a function of fixed parameters of the institutional and informational environment. In any subgame-perfect equilibrium in which the monitor chooses $\pi_i(k)$ and $\pi_i(c)$ such that inequality (1) is not satisfied, no group member will contribute. However, if the monitor chooses $\pi_i(k)$ and $\pi_i(c)$ such that inequality (1) is satisfied, positive contribution levels can be sustained in subgame-perfect equilibria of the stage game. The power of the incentives created by the enforcement regime is clearly maximized in equilibria where the left hand side of (1) is largest, which is achieved when the monitor always punishes given a signal of "kept" ($\pi_i(k) = 1$) and never punishes given a signal of "contributed" ($\pi_i(c) = 0$). We will refer throughout to ($\pi_i(k) = 1$, $\pi_i(c) = 0$) as the "Punish According To Signal" (PATS) strategy for the monitor.

Finally, note that the false positive and false negative error probabilities, q_1 and q_2, are *perfect substitutes* in the best response correspondence of the group members as stated in inequality (1). Holding constant $q_1 + q_2$, whether the error structure is more prone to false positives (which would induce Type I errors – wrongful punishment – by monitors playing the PATS strategy) or to false negatives (which would induce Type II errors – failure to punish the non-compliant – by monitors playing the PATS strategy) should have no effect on a fully rational group member's behavior.

3 Experimental Protocol

We conducted a laboratory experiment to explore the intuitions derived from the rational-choice benchmark. The paper describes data collected during 10 experimental sessions that were carried out in a social science lab at a large American university. Each of the 190 subjects who participated took part in one session only. Subjects interacted anonymously via networked computers. The experiments were programmed and conducted with the software z-Tree (Fischbacher 1999). Participants signed up via a web-based recruitment system that draws on a large, pre-existing pool of potential subjects. (Subjects were not recruited from the authors' courses.) Almost all subjects were undergraduates from around the university. After giving informed consent according to standard human subjects protocols, subjects received written instructions that were subsequently read aloud in order to promote understanding and induce common knowledge of the experimental scenario. No deception was employed in our experiment, in accordance with the long-standing norms of the lab in which the experiment was carried out. Before beginning the experiment itself, subjects took an on-screen quiz that both measured and promoted understanding of the instructions.

At the beginning of each session, subjects were randomly assigned to a group of five people, of which one was randomly assigned to "Role A" while the others were assigned to "Role B." These more neutral labels were employed in lieu of "monitor" and "group member," their referents in the theoretical exposition above.[7] Group and role assignments remained fixed over 20 periods of interaction. However, on the monitors' screens, individual group members in role B were labelled with an "ID number" between 1 and 4 that *was* randomly reassigned every period.

Each period consisted of one play of the Public Goods with Flawed Monitoring Technology stage game, the game-theoretic structure of which was identical to the model described in the previous section. Subjects earned "tokens," convertible into dollars at the end of the experiment (30 tokens = US$1) in an amount determined by the outcome of play. In each period, each group member's contribution decision involved a binary choice between "allocating" an initial endowment of 20 tokens to a "common pot" or "keeping" those 20 tokens "for him- or herself." In all but one of our treatments, the monitor received a per-period endowment ("automatic token supply") of

[7] The Appendix contains a sample set of instructions to subjects, offering a depiction of the way the experiment was framed for participants. In this section, terminology from the experimental scenario is introduced in quotation marks where it differs from the theoretical exposition; for continuity, however, the analysis is presented using the terms introduced earlier.

10 tokens. After receiving a signal about each group member's contribution, the monitor made a series of decisions about whether or not to "reduce" each individual group member's payoffs by 20 tokens. Subjects' overall payoffs were equal to the sum of payoffs from each of the 20 periods, plus a US$5 show-up fee.

As described in the theoretical section, monitors received a (possibly inaccurate) signal about each individual group member's contribution decision. The nature of the process by which signals about contributions were generated varied across three distinct treatments. Three experimental sessions (involving 55 subjects) were exclusively devoted to the *False Positives Treatment*, in which a group member's decision to contribute generated a "Contributed" signal with probability 0.60 and a "Did not Contribute" signal with probability 0.40; a group member's decision not to contribute generated a "Did not Contribute" signal with certainty. Three other experimental sessions (also involving 55 subjects) were exclusively devoted to the *False Negatives Treatment*, in which an individual's decision *not* to contribute generated a "Did not Contribute" signal with probability 0.60 and a "Contributed" signal with probability 0.40; an individual's decision to contribute generated a "Contributed" signal with certainty.

The four remaining experimental sessions (involving 80 subjects) were exclusively devoted to the *Technology Choice Treatment*. In this treatment, before the first period of stage game play, monitors had an initial opportunity to choose "the rules" for their group. If the monitor chose the "Default Rules," he or she would receive perfectly accurate signals about every group member's contribution decision in every period, but would receive no automatic token supply. If the monitor instead chose the "Replacement Rules," play would proceed in precisely the same way as in the False Positives Treatment. Monitors' choices about the rules were made known to their group members, and the rules chosen remained in place for the duration of the experiment. The three treatments differed only in these details about the process by which signals were generated (and the initial choice by the monitor in the "Technology Choice" treatment of whether to forego the automatic token supply). Table 1 lists parameter values for each treatment along with the corresponding notation from the theoretical exposition.

Table 1 about here

At no point did monitors receive further information, beyond the signals described above, about

any *individual* group member's choice. As such, monitors never directly learned about the accuracy of any specific enforcement decision. However, as feedback at the end of each period, the monitor *was* informed of the *overall* level of contributions by the group as a whole. Naturally, this feedback contained information relevant to monitors' assessments of the likely accuracy of their punishment choices. As feedback at the end of each period, each group member was also informed of the overall group contribution level, along with the monitor's enforcement decision relevant to that specific group member himself; no group member ever observed the extent to which other group members were punished (or the contribution decision of any specific other group member).

The specific parameter values in each treatment give rise to behavioral predictions associated with the rational-agent benchmark. In particular, suppose that in a subgame-perfect equilibrium of the stage game, the monitor were to pursue the PATS strategy ($\pi_i(k) = 1$, $\pi_i(c) = 0$) described in the previous section – that is, always punishing given a signal of "kept," and never punishing given a signal of "contributed." In the technology choice treatment, conditional on a monitor foregoing the automatic token supply in return for perfect monitoring capability, group members would contribute in any such (stage-game) subgame-perfect equilibrium, because inequality (1) would be (strictly) satisfied.[8] In the False Positives and False Negatives Treatments, as well as in the Technology Choice Treatment for groups whose monitor did not choose the perfect monitoring capability, the same enforcement strategy would induce indifference in group members, implying that contributions and non-contributions alike could be sustained in subgame-perfect equilibria of the stage game.[9] Notably, under *any other* enforcement strategy, no contributions would take place in these conditions, because (1) would no longer be satisfied. Critically, for any fixed enforcement strategy, group members face identical incentives across these three experimental conditions whether or not the monitor plays the PATS strategy, because $q_1 + q_2 = 0.4$ in each of them.

Of course, the context of interaction within fixed matchings over 20 periods meant that monitors' enforcement decisions could have consequences extending beyond the particular period in which a given decision took place. This feature of the design was important, given our desire to explore how the effectiveness of enforcement regimes varies as behavior plays out over time. Significantly, however, because group members' labels (ID numbers) were randomly reassigned every

[8] Substituting into (1), $1 - 0 > \frac{(.6)20}{(1)20}$, or $1 > .6$.

[9] Substituting into (1), $1 - 0 = \frac{(.6)20}{(.6)20}$, or $1 = 1$.

period, individual group members could not be associated with reputations for contributing (or not contributing) to the public good. Indeed, monitors' decisions about a specific group member in a given period could not be conditioned specifically on that group member's behavior (or individual-specific signals about that behavior) from earlier periods. Nonetheless, monitors could condition enforcement decisions on posterior beliefs about behavior in the past for the group as a whole, a topic we address in greater detail below.

4 Results

4.1 Comparing Behavior in the False Positives and False Negatives Treatments

Group Members' Contribution Decisions: Overall Patterns. We begin by comparing behavior across the False Negatives Treatment (non-contributors sometimes wrongly exculpated) and the False Positives Treatment (contributors sometimes wrongly accused). Figure 1 presents a graphical depiction of period-by-period group contribution rates for both treatments. In period 1, on average, 3.18 out of 4 (79.6%) group members contribute in the False Negatives Treatment, compared to 3.09 out of 4 (77.3%) in the False Positives Treatment, a substantively small and statistically insignificant difference ($p = 0.81$, two-tailed).[10] The Figure shows that contribution rates remain only slightly higher in the False Negatives Treatment than in the False Positives Treatment through early periods (1-6).

FIGURE 1 ABOUT HERE

In later periods (7-20), however, contribution rates diverge more sharply, with contributions under False Negatives consistently and substantially higher than contributions under False Positives. Averaging across all periods of the experiment, the mean contribution rate per group is substantially larger in the False Negatives Treatment (2.64 out of 4, or 66%) than in the False Positives Treatment (1.85 out of 4, or 46.3%). This difference is highly statistically significant ($p = 0.02$, two-tailed). In the False Negatives Treatment, the lowest-performing group had a contribution rate

[10] We employ one-tailed tests when hypotheses are directional in nature, two-tailed tests otherwise. Unless otherwise noted, we report p values from t-tests with unequal variances. In all instances, Wilcoxon rank-sum tests produced highly similar results.

of 40%, while the highest had a rate of 90%. In the False Positives Treatment, by contrast, group averages ranged from 17.5% to 84%.[11]

Why do contribution rates diverge over time between these two treatments despite being similar initially? Before addressing this question directly, it will be useful to describe the monitors' enforcement decisions.

Monitors' Punishment Decisions and Performance: Overall Patterns. Table 2 displays summary information about enforcement decisions made by monitors in the two treatments. The data reveal considerable variation both across and within treatments. Most strikingly, the average rate at which monitors who received an accusatory signal chose to punish the associated individual was 89.2% in the False Negatives Treatment but only 51.6% in the False Positives Treatment, a substantively large and highly statistically significant difference ($p < 0.001$, two-tailed). In the aggregate, monitors in both treatments deviate from the PATS strategy described in section 2, which suggested that the enforcement regime would exert maximum influence over group members if monitors were to punish 100% of the time given a signal of non-contribution. However, the extent of deviation is much smaller in the False Negatives Treatment than the False Positives Treatment. The nature of this across-treatment difference is suggestive of how monitors responded to the possibility of committing Type I errors. In the False Negatives Treatment, an accusatory signal is perfectly informative that a given group member failed to contribute, so a decision to punish conditional on an accusatory signal involves no risk of committing a Type I error. In the False Positives Treatment, an accusatory signal is not perfectly informative, and a decision to punish based on such a signal does carry that risk. In the False Negatives Treatment, seven out of eleven monitors *always* punished given an accusatory signal and an eighth did so 95.0% of the time. All eight of these monitors punish at a higher rate given an accusatory signal than the most aggressive monitor in the False Positives Treatment, who did so 88.0% of the time. The differences across treatments are stark at the other end of the distribution as well. The monitor in the False Negatives Treatment who punishes *least* often given an accusatory signal (54.8% of the time) nonetheless does so at a higher rate than seven out of eleven monitors in the False Positives Treatment.

[11] This across-treatment difference in contribution levels was reflected in subjects' earnings from the experiment. Including the show-up fee, False Negatives monitors earned US$ 25.75 on average while False Negatives group members earned US$20.68; in contrast, False Positives monitors earned US$21.53 on average, while False Positives group members earned US$17.02.

TABLE 2 ABOUT HERE

By contrast, there are no clear differences across treatments in monitor behavior conditional on receiving a signal that a group member contributed. The average monitor punished only 6.6% of the time given an exculpatory signal in the False Negatives Treatment, and 3.3% of the time in the False Positives Treatment. One outlying monitor in the False Positives Treatment (who punished 42.8% of the time upon receiving an exculpatory signal) largely accounts for the modest gap in punishment rates across treatments. The mean rates of punishment are statistically indistinguishable ($p = 0.42$, two-tailed), however. These data imply a much closer correspondence between monitor behavior conditional on an exculpatory signal and the PATS strategy, which suggested that the enforcement regime would exert maximum influence over group members if monitors were to punish 0% of the time given a signal that a group member contributed.

Next, we examine the frequency with which monitors make "punishment errors," and the nature of group members' responses to such punishment errors. A monitor's *Type I error rate* is defined as the rate at which members of her group who actually contribute are nonetheless punished in that period. Likewise, a monitor's *Type II error rate* is the rate at which members of her group who do not contribute are not penalized in that period. Finally, a monitor's *overall error rate* is the fraction of a monitor's enforcement decisions involving either a Type I or Type II error. Panels A and B of Figure 2 plot monitors' Type I and Type II error rates for the False Positives and False Negatives Treatments, respectively. (The number next to each marker is the average contribution rate for an individual monitor's group.)

FIGURE 2 ABOUT HERE

Holding constant group members' contribution levels, a monitor who followed the PATS strategy would commit more Type I errors in the False Positives Treatment than in the False Negatives Treatment. The data indicate that the same is true of monitors in our experiment: the mean Type I error rate was 23.6% under the former treatment, and 7.0% under the latter, a statistically significant disparity ($p = 0.008$, one-tailed). Again holding constant group members' contribution levels, a monitor who followed the PATS strategy would, however, commit *fewer* Type II errors in the False Positives Treatment than in the False Negatives Treatment. Interestingly, the data indicate that this is *not* true of monitors in our experiment: the mean Type II error rate was

50.1% in the False Positives Treatment, and 44.7% in the False Negatives Treatment, a statistically insignificant difference ($p = 0.45$, two-tailed). The proliferation of Type II errors in the False Positives Treatment is attributable to the reluctance of most monitors to trust that treatment's imperfect accusatory signals. Because they often do not punish after observing an accusatory signal, they often fail to punish those who did not contribute. Finally, the mean overall error rate is substantially larger in the False Positives Treatment (38.6%) than in the False Negatives Treatment (19.2%), a statistically significant difference ($p = 0.002$, two-tailed).

Punishment errors and contributions. The above findings suggest a natural hypothesis: differences in contribution rates across treatments correspond to differences in punishment behavior and the resultant errors. A limitation of examining Type I and Type II error rates separately is that each figure is normalized by the frequency of group member contributions. So, for example, a high Type II error rate could correspond to a small number of actual errors if the base contribution rate is very high. Accordingly, panels C and D of Figure 2 plot average group contribution rates against their monitors' *overall* error rates in the False Positives and False Negatives Treatments, respectively. Each panel suggests a strong relationship between these two variables: a simple linear regression indicates a statistically significant negative relationship between them ($p = 0.018$ and $p = 0.003$, two-tailed tests, respectively, for the False Positives and False Negatives Treatments).

Next, we explored the micro-level dynamics of enforcement and compliance by conducting a series of logistic regression analyses, the results of which appear in Table 3. Our aim is to consider the effect of specific punishment errors on contribution behavior in the subsequent period. In the basic specification, the probability that group member i contributes in period t is modeled as a function of his own previous contribution behavior; the previous contribution behavior of his counterpart group members; his own experience of punishment in the previous period, and a period-specific intercept:

$$\begin{aligned}\Pr(\textit{contribute}_{i,t}) \;=\; & \Lambda\big(\alpha_t + \beta_1(\textit{one other contributor})_{i,t-1} + \beta_2(\textit{two other contributors})_{i,t-1} \\ & + \beta_3(\textit{three other contributors})_{i,t-1} + \beta_4 \textit{contribute}_{i,t-1} \\ & + \beta_5 \textit{wrongly punished}_{i,t-1} + \beta_6 \textit{wrongly unpunished}_{i,t-1}\big)\,,\end{aligned}$$

where $\Lambda(\cdot)$ is the cumulative logistic distribution function, *wrongly punished*$_{i,t-1}$ is an indicator variable equal to one if, in the previous period, i contributed but was punished; and *wrongly unpunished*$_{i,t-1}$ is an indicator variable equal to one if, in the previous period, i did not contribute and was not punished.[12] Columns (1) through (3) present results for the False Positives Treatment. The first column reports coefficient estimates from a simple logistic regression with robust standard errors. The second reports estimates from a multilevel random effects logistic regression with group- and individual-specific random effects. The third includes a measure of the historical average contribution rate within a group prior to period t. Because the results are substantively similar, we focus in the discussion that follows on the simple specification from column (1).

Table 3 about here

Our results document a number of significant predictors of contribution behavior.[13] First, an individual's propensity to contribute in the current period is strongly correlated with his or her decision in the prior period and other group members' prior contributions. Holding other group members' prior contributions at two (the median value) and assuming no punishment errors in the previous period, an individual who contributed in the ninth period was 25.6 percentage points more likely to contribute in the tenth period than one who had not ($p = 0.00$).[14] Unless otherwise noted, all remaining marginal effects and simulation results are calculated with these quantities held fixed at the same values. There is also a strong association between other group members' prior contributions and an individual group member's decisions. The model predicts that individuals who did not contribute in the previous period would still do so 25.8% of the time if no other group member had contributed, 43.0% of the time if one other member had contributed, 52.6% of the time if two other members had contributed, and 69.2% of the time if all other group members had contributed. The differences between these figures are all statistically significant at $p < 0.05$ (two-tailed).

Of more immediate relevance is the response of group members to punishment errors. Across model specifications for the False Positives Treatment, both Type I and Type II errors were associated with significantly reduced subsequent compliance. For Type I errors, holding other group

[12] Period 1 is excluded due to the absence of measures of lagged punishment and contribution behavior.

[13] We note that our regression results, as well as our other key qualitative results, are robust to dropping groups one at a time, so that our findings are not an artifact of outlying behavior by a given monitor or group.

[14] Unless otherwise noted, standard errors for marginal effects were calculated using the Delta method.

members' prior contributions at two, an individual who was wrongly punished in the ninth period was 12.8 percentage points less likely to contribute in the tenth period than an individual who had contributed without being punished ($p = 0.03$, two-tailed). Type II errors were estimated to be associated with a similar reduction in compliance; holding other variables fixed at the same values, an individual who "got away" with not contributing in the ninth period was 12.0 percentage points less likely to contribute in the tenth period than an individual who had been punished for failing to contribute. In none of our specifications was the effect of being wrongly punished statistically distinguishable from that of being wrongly unpunished.[15]

Columns (4) through (6) present parallel results for the False Negatives Treatment. Per the column (4) specification, holding other group members' prior contributions at two and assuming no punishment errors in the previous period, an individual who contributed in the ninth period was 20.7 percentage points more likely to contribute in the tenth period than one who had not ($p < 0.001$). Interestingly, the relationship between individual contributions and other group members' prior contributions was weaker in the False Negatives Treatment than in the False Positives Treatment. Given that an individual did not previously contribute, the model predicts that individuals contribute 42.8% of the time if no other group member had contributed, 39.7% of the time if one other member had contributed, 50.1% of the time if two other members had contributed, and 61.1% of the time if all other group members had contributed. Only the last figure is statistically distinguishable from the other three.

Turning to the relationship between punishment errors and contribution levels in the False Negatives Treatment, we again estimate a strong effect of both Type I and Type II errors. For Type I errors, the column (4) specification suggests that an individual who was wrongly punished was 21.8 percentage points less likely to contribute in the next period, relative to the baseline 71.6% contribution rate for an individual who had contributed without being punished ($p = 0.036$, two-tailed). The specifications in columns (5) and (6) estimate the effect to be of similar magnitude, although the relevant coefficients are not statistically significant; given the rarity of Type I errors in the False Negatives Treatment, the coefficients are very imprecisely estimated, so this is not altogether surprising. The estimates for Type II errors, again, were similar; an individual who "got

[15]The *lowest* two-tailed p-value associated with the test of the hypothesis that $\beta_5 = \beta_6$ was 0.326, from specification (2).

away" with not contributing was 22.8 percentage points less likely to contribute in the following period ($p < 0.001$). This result was more statistically robust across specifications. Finally, as in the False Positives Treatment, in none of the specifications was the effect of Type I errors statistically distinguishable from the effect of Type II errors.

In light of the above, an interesting question is whether group members respond differently to being wrongfully punished, or to being wrongfully unpunished, in the different contexts of the False Positives and False Negatives Treatments. The specifications in column (7) addresses this question by pooling observations across treatments and interacting the effects of punishment errors with a False Negatives Treatment indicator. To test for significant differences in the response to punishment errors across treatments, we examine the interaction between the False Negatives Treatment indicator and each punishment error variable. In both cases, we cannot reject the null hypothesis that the response to each type of punishment error is the same across treatments ($p = 0.659$ and $p = 0.407$, respectively, for the incremental effects of *wrongly punished*$_{i,t-1}$ and *wrongly unpunished*$_{i,t-1}$).

Anticipation of Behavioral Response in the False Positives Treatment. The foregoing analysis permits us to infer that in both the False Positives and False Negatives Treatments, both Type I and Type II punishment errors have significant negative consequences for subsequent compliance behavior. Given imperfect information about group members' choices, such deleterious effects may significantly hamper monitors' efficacy in promoting group contributions. Importantly, though, this problem is more insidious in one of our treatments than it is in the other. First, consider the False Negatives Treatment. In a setting where compliance levels are high, monitors will receive few incorrect signals about group member behavior (because false negatives can take place only when a player does not contribute) and, therefore, monitors may be able to carry out enforcement without making many punishment errors at all. Group members may in turn be deterred, at least partially, from non-compliance by the threat of such effective enforcement. In contrast, the prospect of making errors in enforcement poses a stiffer challenge for monitors in the False Positives Treatment. In a setting where compliance levels are high, monitors will receive *many* incorrect signals about group member behavior – thereby raising the likelihood that monitors will err in their enforcement decisions. Given the structure of information in the False Positives Treatment, monitors receive *more* incorrect signals about group member choices when overall compliance rates

are *higher*.

This observation, along with our earlier results on the deleterious effects of punishment errors, suggests a natural question: given group members' behavioral responses to punishment, when *should* a monitor punish a group member about whom she has received an accusatory signal? Put differently, over what range of prior beliefs about a group member's underlying probability of contributing should a monitor act on an accusatory signal, and over what range of prior beliefs should a monitor ignore such a signal? The logic of the previous paragraph suggests that, in certain circumstances, a monitor may indeed be better off not punishing given an accusatory signal. If she believes the baseline rate of contributions to be very high, then a given accusatory signal is very likely to be in error, and punishment is likely to be counterproductive. On the other hand, if she believes the baseline rate of contributions to be very low, then accusatory signals are very likely *not* to be in error, and a decision *not* to punish could well be the counterproductive choice.

Let $\theta \in [0,1]$ represent a monitor's prior belief that a given group member i will contribute in a given period. Then by Bayes' Rule, upon receipt of an accusatory signal, a Bayesian monitor would have posterior belief

$$\Pr(C_{i,t} = 1 | s_i = k) = \frac{0.4\theta}{1 - 0.6\theta}$$

that player i in fact contributed. Let $p^{wp}_{i,t+1}$ be the probability that group member i contributes in period $t+1$ given that he was subject to a Type I error (that is, was wrongfully punished) in period t; let $p^{cp}_{i,t+1}$ be the probability that i contributes in period $t+1$ given that he contributed and was not punished in period t; let $p^{wn}_{i,t+1}$ be the probability that i contributes in period $t+1$ given that he was subject to a Type II error (that is, did not contribute and was not punished) in period t; and let $p^{cn}_{i,t+1}$ be the probability that i contributes given that he did not contribute and was punished in period t. Recalling that ry_g is the marginal return to the monitor of inducing compliance by an individual group member, the expected utility to the montior in period $t+1$ from punishing i, and from not punishing i, in period t are respectively given by

$$\begin{aligned} E[u_{m,t+1}(P_i = 1 | s_{i,t} = k)] &= \left(\frac{0.4\theta p^{wp}_{i,t+1}}{1 - 0.6\theta} + \frac{(1-\theta) p^{cp}_{i,t+1}}{1 - 0.6\theta} \right) r y_g \\ E[u_{m,t+1}(P_i = 0 | s_{i,t} = k)] &= \left(\frac{0.4\theta p^{cn}_{i,t+1}}{1 - 0.6\theta} + \frac{(1-\theta) p^{wn}_{i,t+1}}{1 - 0.6\theta} \right) r y_g. \end{aligned}$$

Comparing these expressions, the monitor is better off punishing if and only if $0.4\theta(p^{wp}_{i,t+1} - p^{cn}_{i,t+1}) + (1-\theta)(p^{cp}_{i,t+1} - p^{wn}_{i,t+1}) > 0$. Substituting empirically estimated values of $p^{wp}_{i,t+1}$, $p^{cn}_{i,t+1}$, $p^{cp}_{i,t+1}$, and $p^{wn}_{i,t+1}$ from our statistical analysis of contribution decisions and accounting for uncertainty in their estimation via stochastic simulation, we can estimate the extent to which punishment is helpful or harmful as θ varies.[16]

Figure 3 displays the results of such a simulation analysis for the False Positives Treatment. The Figure displays the simulated *probability* that punishing, given an accusatory signal, improves compliance in the subsequent period as prior beliefs about compliance vary. (For example, a monitor may form beliefs for period t based on observed group contributions in earlier periods.) The lines correspond to simulations run using parameter values obtained from specifications (1) and (2) from Table 3.[17] As expected, the probability that punishment is beneficial is estimated to be high when θ is low, but to decrease monotonically as θ increases. Naturally, the point at which that probability drops below 50% is of particular interest. Using the results from specification (1), the logit with robust standard errors, a monitor receiving an accusatory signal will be better off punishing the alleged non-complier only if she has a (correct) prior belief that the baseline probability of contribution is less than 70%. Using the results from specification (2) instead, the logit with multilevel random effects, the same is true only under the more restrictive condition that the baseline probability of contribution is less than 55%.

FIGURE 3 ABOUT HERE

4.2 Technology Choice Treatment

We now turn our attention to the four experimental sessions devoted to the Technology Choice Treatment. Before proceeding to the results, we first note that given the experiment's parameters, under the "default" rules (perfect information about contributions) the monitor should be able to induce perfect compliance by implementing the PATS strategy described in Section 2. Perfect compliance is worth 640 tokens to the monitor (0.4 × 20 tokens × 4 group members × 20 periods).

[16] We emphasize that this approach analyzes a monitor's best response "in the data," that is, given overall observed group member behavior; it is, of course, not an equilibrium-based analysis. We also note that the analysis uses estimates of behavior in the tenth period of play – that is, halfway through an experimental session – except that, for consistency, the number of "other" group members contributing is determined in our simulation by carrying out multiple draws from the distribution defined by θ.

[17] The plot from specification (3) is nearly identical to that of (1), and is thus omitted for clarity.

However, the monitor must forego an automatic token supply worth 200 tokens to retain the perfect technology. Consequently, a rational monitor who believed perfect compliance was possible under the default rules would choose the "replacement" rules (with false-positive-prone technology) only if she expected group members to contribute *at least* 68.75% (440/640) of the time under the replacement rules.

In the four sessions, 62.5% of monitors (10/16) chose the "replacement rules" over the "default rules," trading a perfect monitoring technology without an automatic token supply for the strategic equivalent of the False Positives Treatment described above. Monitors choosing to remain under the default rules came close to implementing the PATS strategy; five of six punished 100% of the time given an accusatory signal, while the sixth punished 93.3% of the time. Consequently, total punishment error rates were very low under the default rules – 2.9%, compared to 34.4% under the replacement rules. Not surprisingly, contributions were far higher on average under the default rules: the mean group-level contribution rate was 88.5% under the default rules, but only 44.6% under the replacement rules ($p = 0.003$, two-tailed).[18]

Only two of the ten monitors who chose the replacement rules were able to surpass the 68.75% benchmark described above. A third monitor exactly achieved that level of contribution in her group. If we employ the observed average 88.5% contribution figure as a measure of expected contribution rates under the default rules (rather than 100%), the remaining seven monitors *still* failed to meet the revised benchmark (57.3%) that would justify switching to the replacement rules.[19]

Because our interest is in the dynamics of enforcement and compliance when signals about behavior are imperfect, our focus in the remainder of this section is behavior in groups operating under the Replacement Rules. For clarity of exposition, we refer to this as the Technology Choice False Positives condition. Figure 4 depicts group average contribution rates in that condition by period, for comparison juxtaposed against the analogous results from the "exogenous" False Positives Treatment discussed previously (and displayed in Figure 1). Strikingly, the rate of first-period group contributions is much higher in the exogenous False Positives Treatment (3.09 out of

[18]Including the show-up fee, under the default rules, Technology Choice monitors earned US$23.89 on average while Technology Choice group members earned US$23.56. By contrast, under the replacement rules, Technology Choice monitors earned US$21.19 on average while Technology Choice group members earned US$15.80.

[19]57.3% is arrived at by solving the following equation for x: (0.88 - x) $\times$ 0.4 $\times$ 20 tokens $\times$ 4 group members $\times$ 20 periods = 200.

4 group members, or 77.3% on average) than in the corresponding Technology Choice data (1.80 group members, or 45.0%), and this difference is highly statistically significant ($p = 0.008$, two-tailed). This is true even though both sets of subjects were, at that point, interacting in the context of *identical* rules. However, as is evident from Figure 4, this across-treatment difference vanishes quickly. Indeed, the average group contribution rates across all periods were nearly identical in each treatment: 46.3% in the exogenous False Positives treatment, and 44.6% in the Technological Choice False Positives condition ($p = 0.883$, two-tailed).[20]

FIGURE 4 ABOUT HERE

What accounts for the initial divergence, but subsequent convergence, of group contribution rates in the two false positives scenarios? One possible answer is that monitors in the Technology Choice False Positives condition manage to overcome group members' initial inclinations not to contribute by punishing more frequently given accusatory signals, both in the first period and after. Table 4 displays summary data on monitors' punishment choices in that condition. Consistent with this intuition, conditional on observing an accusatory signal, monitors on average in the first period punished 47.5% of the time in the Technology Choice treatment and 30.0% of the time in the exogenous False Positives Treatment. However, this difference is not statistically significant at conventional levels ($p = 0.42$, two-tailed), perhaps unsurprising given the limited amount of data available for a single period. Across all periods, the mean punishment rate conditional on receiving an accusatory signal was also higher under Technology Choice False Positives than in the exogenous False Positives treatment: 64.9% versus 51.6%, a difference that narrowly misses statistical significance at conventional levels ($p = 0.13$, two-tailed).

TABLE 4 ABOUT HERE

Interestingly, over all 20 periods the total rate of punishment errors is quite similar across these settings: 38.7% in the exogenous False Positives Treatment, and 34.4% under the Technology Choice False Positives condition ($p = 0.38$, two-tailed). Panel (B) of Figure 5 plots individual monitor error rates and average contribution rates for groups who operated under the Technology Choice

[20] Period-by-period comparisons after the first period likewise show no statistically significant across-treatment differences in contribution rates.

False Positives condition. The pattern is very similar to the one revealed in Figure 2. We continue to observe a strong negative relationship between total punishment errors and contributions.

FIGURE 5 ABOUT HERE

We supplement this graphical depiction with statistical analyses of the relationship between punishments received (or not received) by a group member and that group member's contribution behavior. These analyses are shown in Table 5. Columns (1)-(3) indicate that, as in the False Positives Treatment, Type I (wrongful punishment) and Type II (wrongful non-punishment) errors both have statistically significant and substantively important negative effects on subsequent contributions in the Technology Choice False Positives condition. Per the column (1) specification, a group member who was wrongly punished in the ninth period was estimated to be 14.8 percentage points less likely to contribute in the tenth period than one who had contributed without being punished ($p = 0.035$, two-tailed; the baseline probability was 71.9%). Likewise, a group member who had *not* contributed in the ninth period and was *not* punished was 19.1 percentage points less likely to contribute in the tenth period than one who had been punished following a non-contribution ($p = 0.003$, two-tailed; from a baseline probability of 32.4%). While the coefficient estimate for getting away with non-contribution is larger than that for being wrongfully punished, the two effects are not statistically distinguishable ($p = 0.24$).

TABLE 5 ABOUT HERE

In columns (4)-(7), we compare the effects of wrongful punishment and wrongful non-punishment under exogenous False Positives and Technology Choice False Positives, constraining the effect of prior individual and group contributions to be identical across treatments. Across specifications, group members who fail to contribute, and are not punished, decrease their contributions by a significantly larger amount in the Technology Choice False Positives condition than in the False Positives Treatment, and this difference is statistically significant (the maximum p-value across specifications is 0.04). However, across specifications, there is no statistically significant difference across treatments in the effect of wrongful punishment on group member behavior (the minimum p-value across specifications is 0.612). It therefore appears that "getting away with cheating" increases subsequent noncompliance more in the Technology Choice False Positives condition than in the exogenous False Positives treatment.

Finally, using the coefficient estimates from the Technology Choice False Positives regressions, we again conducted the simulation analysis introduced in the previous section. Specifically, we explored the extent to which punishing conditional on an accusatory signal would be beneficial for monitors who had correct prior beliefs about individual group members' likelihoods of contribution. Figure 6 displays the probability that punishment is beneficial as a function of such beliefs. At first glance, the figure looks very similar to that derived for the (exogenous) False Positives Treatment (Figure 3). However, there is an important difference: punishment conditional on an accusatory signal is more likely to be beneficial for a broader range of beliefs under the Technology Choice False Positives condition. Using the results from specification (1), the logit with robust standard errors, a monitor receiving an accusatory signal will be better off punishing the alleged non-complier if she has a (correct) prior belief that the baseline probability of contribution is less than 82% (compared with 70% for the exogenous False Positives Treatment). Using the results from specification (2) instead, the logit with multilevel random effects, the same is true only under the more restrictive condition that the baseline probability of contribution is less than 69% (compared with 55% for the exogenous False Positives Treatment). These differences are substantively significant, corresponding to an additional one-half of a contribution per period in groups of four for which punishing according to signal would remain beneficial in the Technology Choice False Positives Condition but not in the exogenous False Positives Treatment.

FIGURE 6 ABOUT HERE

5 Discussion

In this section, we discuss three implications of our analysis. First, we consider the ramifications of the observed differences in behavior across the False Positives and False Negatives Treatments. Second, we consider how the comparison of the exogenous False Positives Treatment and the Technological Choice False Positives condition can help to shed light on the behavioral types of monitors. Finally, we discuss the implications of our findings under Technological Choice False Positives for theories of government legitimacy.

5.1 The False Positives Trap

The comparison between the False Negatives (inaccurate exculpatory signals) and False Positives (inaccurate accusatory signals) Treatments suggests an important feature of governing institutions prone to false positives and, thus, Type I errors (wrongful punishment). In the stage-game rational-agent benchmark, the probabilities of false positive signals and false negative signals are perfect substitutes in group members' best response correspondences. This does not mean, however, that Type I and Type II errors have symmetric *behavioral* effects as monitors and group members interact with one another over time.

Consider an environment in which the threat of punishment encourages compliance, but the compliant are sometimes erroneously punished. In the context of the rational-agent benchmark, subgame-perfect equilibria exist in which group members would simply accept the occurrence of such errors as a (possibly unavoidable) feature of the enforcement regime, and contribute anyway. As our micro-level analysis of contribution behavior demonstrated, however, group members in our laboratory experiment did not accept wrongful punishment with such equanimity.[21] Instead, on average they reacted negatively both to Type I and to Type II punishment errors. This behavioral reaction put False Positives Treatment monitors in a difficult position: punish given an accusatory signal, in which case you may err (by punishing a contributor) and discourage contribution in the subsequent round; or don't punish given an accusatory signal, in which case you may *still* err (by failing to punish a non-contributor), also discouraging contribution in the subsequent round. As our simulations demonstrate, weighing the pros and cons to determine whether, in the face of these effects, punishing is more likely to be harmful or beneficial is far from a trivial exercise. Facing this difficulty, most monitors in the False Positives Treatment failed to be resolute; the consequence was to undermine the efficacy of the enforcement regime and further discourage compliance. This is the "false positives trap." The logic of the false positives trap suggests that sustaining relatively high levels of contribution may be particularly difficult. When compliance is fairly high on average, many false positive signals will be generated; as such, monitors must choose either to punish, thereby

[21] In addition to our earlier results, we note the following anecdote from the data. In 12 instances during the False Positives Treatment sessions, a group member was wrongfully punished during a period in which his group experienced a contribution level of 100%. Even in such a setting, in which one might think there is an especially high incentive to continue contributing to a group that is functioning well, 41.7% (5/12) of these group members chose not to contribute in the subsequent period.

committing many Type I errors, or to abdicate their punishment role, vitiating the power of the enforcement regime. Either course of action is likely to lead to lower contributions over time.

Compare this situation to one in which the threat of punishment is effective, but in which the non-compliant might sometimes be "let off the hook." Again, in the context of the rational-agent benchmark, subgame-perfect equilibria exist in which group members would choose to contribute to the public good. However, in this False Negatives setting, no equanimity would be required from group members because in such an equilibrium *punishment errors never occur* – monitors would exercise their punishment capacity only off the equilibrium path of play. As such, a monitor under False Negatives does not face the same tradeoffs as a monitor under False Positives; there are no comparable costs to being resolute in the face of an accusatory signal. Accordingly, there is no comparable "false negatives trap," and sustaining relatively high levels of contribution is less difficult than under False Positives. After all, when compliance is fairly high on average, few erroneous signals will be received by the monitor under False Negatives, so punishment error rates will tend to be quite low.

5.2 Selection Effects

A second implication of our findings concerns variation in the types of individuals assigned enforcement responsibilities. Our benchmark rational agent model did not allow for heterogeneity in, *inter alia*, the skill, resoluteness, vengefulness, or greediness of the monitor: all individuals occupying that role in the laboratory had access to the same information and were subject to the same incentives. However, we documented substantial differences *within treatments* in the punishment behavior of monitors. These differences persisted in spite of the fact that they had significant implications for contribution behavior, and, ultimately, the payoffs of the monitors themselves. The existence of this variation in the exogenous False Positives and False Negatives Treatments is, in itself, evidence that individuals may be heterogeneous in a number of unmodeled respects, potentially including the extent to which they are squeamish about punishing others given the possibility of error, and the extent to which they comprehend the logic of deterrence.

In this respect, one may interpret the Technology Choice Treatment as giving individuals assigned to the monitor role the opportunity to sort by type. Indeed, we observe differences in the punishment behavior of monitors in the exogenous False Positives Treatment when compared

with those in the Technology Choice False Positives condition. However, it is not immediately clear whether we can ascertain the precise nature of any sorting that does occur. Monitors who maintained the "default rules" (perfect information), for example, may have done so because they anticipated that the value of enhanced deterrence would more than make up for the forgone value of the automatic token supply (as, empirically, it appears to have done). Other monitors may instead have maintained the "default rules" because they did not wish to confront the possibility of "unfairly" punishing group members who did contribute. Monitors who chose the "default rules" for the former reason, given their degree of insight, may well have been effective as monitors under the "replacement rules"; monitors who chose the "default rules" for the latter reason, given their reluctance to exercise authority in an uncertain world, would likely have been ineffective as monitors under the "replacement rules." As such, it is ambiguous whether those individuals would have induced higher or lower contributions relative to the average, had they been assigned to a treatment in which the rules were exogenous.

At the same time, it is also possible that differences in punishment behavior between the Technology Choice False Positives condition and the exogenous False Positives Treatment may be a consequence not of type selection, but of differences in the circumstances in which both monitors and group members find themselves at the start of play. We now turn our attention to this possibility.

5.3 Legitimacy and the Selection of Rules

Our findings comparing behavior in the (exogenous) False Positives Treatment and the Technology Choice False Positives condition have implications for understanding how citizens view agents with coercive authority, and how in turn those agents react to citizens' perceptions. While both conditions were conducted under an identical error-prone signaling technology, this technology was the result of monitor choice in the Technology Choice Treatment, but was simply an exogenous "fact of life" in the False Positives Treatment. We observed substantial differences across these conditions in *initial* contribution behavior, but no significant differences over the long run. In the False Positives Treatment, where the error-prone technology was exogenous, a larger proportion of group members initially contributed than in the Technology Choice False Positives condition, in which the monitor had chosen to forego a perfect-information signaling technology that had been

feasible *ex ante*. Critically, these initial across-treatment differences were observed *before* any group members had experience with *actual* punishment decisions by their monitor.

At least two interpretations of these results are possible. According to one interpretation, the first-period difference in contributions reflects differences in group members' impressions of the fairness, and through this ultimately the legitimacy, of their monitor (with enforcement technology itself held constant).[22] Such differences could naturally stem from the circumstances under which the institution was adopted, and in particular the fact that, under Technology Choice, it was possible that a superior institution might have been adopted. This interpretation is consistent with the views expressed by a number of group members in the Technology Choice Treatment in debriefing questionnaires that were administered at the conclusion of each experimental session. Recall that in the experimental frame presented to subjects,"Role A" referred to monitors and "Role B" to group members. One subject, whose group interacted under the replacement rules, said that (s)he "didn't agree with [the choice of rules] since the people in Role B would be treated unfairly." Others, interacting under the default rules, said that their monitor's choice "showed that they were a fair person who was looking for everyone's gain," or "meant A was probably a fair person." Another stated, "I was glad they chose that, so the people who didn't put in their coins could be fairly punished."

A second interpretation reflects a subtly different notion of legitimacy: by taking the sure payoff at the expense of accuracy, monitors choosing the replacement rules may have created suspicion among group members either about the monitor's motivations or the monitor's innate competence as an agent in improving group welfare. Further, the technology choice was *commonly known* to all group members – they not only observed the choice, they also knew that other group members observed it, and that those other group members in turn knew that they themselves had observed it. In announcing that she was willing to sacrifice an institution conducive to encouraging group contributions, the monitor may have set the tone for group interactions, unwittingly encouraging a reaction through which individual group members, anticipating a future of poor group performance, decided to begin by withholding contributions themselves. Such thinking is evident in debriefing responses composed by a number of group members who played under the replacement rules:

[22]We rely on the following definition of legitimacy from Tyler and Huo: "[L]egitimacy is the property that a rule or an authority has when others feel obligated to defer voluntarily to that rule or authority" (2002: 101).

- "... was very selfish because Role A was thinking more about his/her own benefit rather than the benefit of the entire group"
- "The Role A person would've in fact made much more money if this person had chosen the default rules and was able to punish Role B people severely in the first few rounds until every Role B person was coerced to allocate to the common pot. Role A is an idiot."
- "I think that the default rules would have been more conducive to group cooperation, but instead it set up the experiment on a more negative level, making people feel like they had to follow suit by making decisions based on what would be the most directly beneficial to them. It reminded me of an 'every man for himself' situation."

The last comment is particularly striking in its recognition of the initial atmosphere that the technology choice may have created.

By the same token, contribution levels under the Replacement Rules quickly came to approximate those in the exogenous False Positives Treatment. Does this mean that group members simply got past their initial distrust of the monitor? That is one possibility. Another, however, is that faced with initially low contribution levels, monitors who chose the replacement rules had stronger incentives to approximate the PATS strategy. (As we note above, it may also be the case that monitors who selected replacement rules were simply less squeamish about committing Type I errors.) That punishment behavior may have, in turn, induced higher levels of subsequent contributions. In fact, as described above, we did observe higher rates of punishment conditional on accusatory signals under Technology Choice False Positives than in the False Positives Treatment, the difference falling just a bit short of statistical significance at conventional levels.

Further evidence for the importance of more aggressive punishment behavior in the Technology Choice False Positives condition may be found in Figure 7. Recall from Section 2 that in the rational-agent benchmark, the power of incentives to comply is maximized when the monitor always punishes given an accusatory signal, and never punishes given an exculpatory one. The figure plots differences in average conditional punishment rates for each monitor (the empirical analog of the left side of inequality (1)) against average group contributions for the three conditions with imperfect monitoring. A version of the rational-agent model with random preference shocks would anticipate a positive relationship between these two variables. For all three conditions, the slopes are positive

and either statistically or nearly-statistically distinguishable from zero.[23] The figure also indicates that the regression line associated with the Technology Choice False Positives Condition has a steeper slope than in the other treatments, which is confirmed in a pooled regression analysis with separate slopes for each treatment.[24] We note that it is at the low end of the empirical measure of incentive power where differences in behavior across treatments are most pronounced, consistent with the claim that aggressive enforcement is most needed when group members are otherwise disinclined to contribute.

FIGURE 7 ABOUT HERE

Finally, this intuition was further borne out by our simulation results, which show that punishment conditional on an accusatory signal was beneficial over a broader range of beliefs about contribution behavior in the Technology Choice False Positives condition than in the exogenous False Positives Treatment.

The combination of these pieces of evidence suggests that it was not the return of legitimacy (e.g., an inherent willingness to support an institution in the absence of enforcement), but rather the combination of initial distrust *and* subsequent resolute punishment that restored contributions to levels comparable to those in the exogenous False Positives treatment.

An open question for future research concerns the extent to which monitors, in different institutional settings, may be able to recover in practical terms from an episode that depletes their legitimacy. We note that, despite the considerable rates of non-contribution in all of our treatments, monitors in our study were actually quite powerful: a choice to punish a group member could reduce that group member's payoffs by an amount equal to his *entire* initial endowment. In other enforcement settings, in which monitors carry smaller sticks, they will have a harder time imposing substantial costs on those who disobey. It is reasonable to hypothesize that a loss of legitimacy may have greater and longer-term negative consequences for expected levels of compliance in those circumstances.

[23] In the False Positives Treatment the coefficient on *Observed Incentives* is 0.29 with a $p = 0.11$ (one-tailed test). In the False Negatives Treatment, the slope is 0.34 with $p = 0.003$. In the Technology Choice False Positives condition it is 0.77 with a $p = 0.00$.

[24] In a pooled regression using all 32 data points in the figure of group contribution rate on separate indicator variables for each treatment, *observed incentives*, and the interaction of *observed incentives* and the Technology Choice False Positives indicator, the coefficient on the interaction term is positive and statistically significant with a p-value of 0.004 (two-tailed test).

6 Conclusion

How does imperfect monitoring, in the form of false positive (incorrectly accusing the innocent) and false negative (erroneously exculpating the guilty) signals, affect the dynamics of enforcement and compliance? To answer this question, we implemented a laboratory experiment that varied the *nature* of imperfections in monitoring across treatments. In contrast to expectations derived from a rational-choice benchmark, we find that compliance in the form of contributions to a public good is substantially lower in an environment characterized by a given rate of false positive signals about behavior, relative to an environment characterized by an equal rate of false negative signals.

This pattern appears to arise for two reasons, which reflect important and previously understudied behavioral dynamics. First, group members' propensities to contribute to a public good are diminished both in response to having been punished despite contributing (Type I errors) and to having escaped punishment despite not contributing (Type II errors). The effects of each type of punishment error are statistically indistinguishable in our data. Second, enforcers appear reluctant to punish those accused of noncompliance when such accusations may be false. However, this reluctance leads them to fail to punish the *guilty*, therefore contributing to a general erosion of the enforcement regime's effective deterrence. We refer to the logic underlying the dilemma faced by enforcers under such circumstances as a "false positives trap." In a generally compliant population, enforcers in our False Positives Treatment receive a substantial number of incorrect signals about behavior, and are caught between the possibility of committing Type I and Type II errors, both of which exert downward pressure on compliance. Of critical importance, there is no corresponding "false negatives trap."

Our study also explores how compliance and enforcement are affected by the *origins* of an enforcement regime. We found that group members were initially much less inclined to contribute to the public good when the flawed False Positives technology was chosen by the monitor, relative to a situation in which the *same* False Positives technology was simply an exogenous fact of life. We interpret this finding as reflecting a difference in the extent to which the monitoring regime was considered legitimate by group members. The False Positives monitoring regime may have been considered less legitimate when selected by the enforcer, either because group members found the enforcer's choice of regimes to be unfair, or because group members anticipated that behavior would

be more self-interested in the aftermath of the enforcer's decision. However, in our experiment, aggressive enforcement was able to close the gap in contributions between the two False Positives conditions. Despite this, we also found that failures to punish non-contributors had more deleterious effects on subsequent compliance when the False Positives regime was chosen by the monitor, suggesting at least some enduring (if surmountable) effects of group members' initial judgments about legitimacy.

Overall, our results have important implications for understanding the dynamics of enforcement and compliance under imperfect information. Given the observed behavior of monitors and group members, our findings suggest that institutions that minimize the likelihood of false positive signals about citizens may be more desirable not only for fairness reasons, but for efficiency reasons as well. If, however, false positive signals are an unavoidable fact of life in some settings, our results stress the value of a relatively sophisticated punishment strategy that takes into account baseline rates of citizen compliance in light of observed behavioral responses to enforcement errors.

Our work points to a number of important areas for future research. In our study, both enforcers and group members shared an interest in maximizing production of the public good. However, in many settings of interest, enforcers extract resources from group members using their coercive abilities, a dynamic omitted from our framework. One might consider the relative weight of an enforcer's motivations and the accuracy of her determinations in affecting overall compliance. A second area for future research is to understand in more detail *how* and *why* a given institution for enforcement comes to be seen as legitimate or illegitimate. The answers to such questions will have important positive as well as normative consequences.

References

Alm, James, Gary H. McClelland, and William D. Schulze. 1999. "Changing the Social Norm of Tax Compliance by Voting," *KYKLOS*, 52(2): 141-71.

Blackstone, William. 1769. *Commentaries on the Laws of England.* Oxford: Clarendon Press.

Bochet, Olivier, Talbot Page, and Louis Putterman. 2006. "Communication and Punishment in Voluntary Contribution Experiments," *Journal of Economic Behavior & Organization* 60(1): 11-26.

Canes-Wrone, Brandice, and Kenneth Shotts. 2007. "When do Elections Encourage Ideological Rigidity?" *American Political Science Review* 101: 273-288.

Dal Bo, Pedro, Andrew Foster, and Louis Putterman. 2007. "Institutions and Behavior: Experimental Evidence on the Effects of Democracy." Brown University Working Paper.

Fehr, Ernst and Simon Gaechter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90(4): 980-994.

Feld, P. Lars and Bruno S. Frey. 2007. "Tax Compliance as the Result of a Psychological Tax Contract," *Law and Policy*, 29(1 Jan.): 102-120.

Fischbacher, Urs. 1999. "z-Tree Zurich Toolbox for Readymade Economic Experiments – Experimenters Manual." Working Paper Nr. 21, Institute for Empirical Research in Economics, University of Zurich.

Gordon, Sanford C., and Gregory A. Huber. 2002. "Citizen Oversight and the Electoral Incentives of Criminal Prosecutors." *American Journal of Political Science* 46: 334-351.

Hardin, Russell. *Trust.* Cambridge, UK: Polity.

Kaplow, Louis, and Steven Shavell. 1994. "Accuracy in the Determination of Liability." *Journal of Law and Economics* 37: 1-15.

Ledyard, John. 1995. "Public Goods." In *The Handbook of Experimental Economics*, John H. Kagel and Alvin E. Roth, eds. Princeton: Princeton University Press.

Png, I.P.L. 1986. "Optimal Subsidies and Damages in the Presence of Judicial Error." *International Review of Law and Economics* 6: 101-105.

Polinsky, A. Mitchell, and Steven Shavell. 2000. "The Economic Theory of Public Enforcement of Law." *Journal of Economic Literature* 38: 45-76.

Scholz, John T. and Mark Lubell. 1998. "Adaptive Political Attitudes: Duty, Trust, and Fear as Monitors of Tax Policy," *American Journal of Political Science*, 42(3 Jul.): 903-920.

Scott, Wilbur J. and Harold G. Grasmick. 1981. "Deterrence and Income Tax Cheating: Testing Interaction Hypotheses in Utilitarian Theories," *The Journal of Applied Behavioral Science*, 17(3): 395-408.

Scholz, John T. and Mark Lubell. 1998. "Trust and Taxpaying: Testing the Heuristic Approach to Collective Action," *American Journal of Political Science*, 42(2 Apr.): 398-417.

Scholz, John T. and Neil Pinney. 1995. "Duty, Fear, and Tax Compliance: The Heuristic Basis of Citizenship Behavior," *American Journal of Political Science*, 39(2 May): 490-512.

Tyler, Tom R. and Yuen J. Huo. 2002. *Trust in the Law.* New York: Russell Sage Foundation.

Tyler, Tom R. 1990. *Why people obey the law.* New Haven: Yale University Press.

Table 1: Parameter Values Across Treatments

		Treatment			
		Exogenous		Technology Choice	
Parameter	*Description*	False Pos.	False Neg.	Default Rules Perfect	Repl. Rules False Pos.
y_g	group member endowment	20	20	20	20
y_m	monitor endowment	10	10	0	10
r	marginal return	0.4	0.4	0.4	0.4
p	penalty	20	20	20	20
q_1	Pr(false positive)	0.4	0	0	0.4
q_2	Pr(false negative)	0	0.4	0	0

Note: Endowments are awarded in each period.

Table 2: Comparison of Group-Level Punishment and Contribution Rates in False Positives and False Negatives Treatments

Punishment Rate Given Accusatory Signal	Punishment Rate Given Exculpatory Signal	Difference in Punishment Rates	Contribution Rate
False Positives Treatment			
0.880	0.033	0.847	0.538
0.785	0.000	0.785	0.425
0.627	0.000	0.627	0.675
0.595	0.000	0.595	0.838
0.507	0.000	0.507	0.175
0.484	0.063	0.421	0.463
0.493	0.077	0.416	0.238
0.424	0.071	0.353	0.338
0.302	0.000	0.302	0.713
0.258	0.000	0.258	0.350
0.317	0.118	0.199	0.338
False Negatives Treatment			
1.000	0.000	1.000	0.763
1.000	0.000	1.000	0.613
1.000	0.016	0.984	0.650
1.000	0.043	0.957	0.775
1.000	0.063	0.937	0.700
1.000	0.067	0.933	0.625
0.950	0.017	0.933	0.563
1.000	0.092	0.908	0.900
0.714	0.000	0.714	0.863
0.606	0.000	0.606	0.400
0.548	0.429	0.119	0.413

Notes: Each entry is a single group observed over twenty periods. Sorted within treatment by difference in punishment rates.

Table 3: Predicting the decision to contribute in the False Positives and False Negatives Treatments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	False Positives			False Negatives			Pooled
	Robust SE	Random group and individual effects	With historical average, robust SE	Robust SE	Random group and individual effects	With historical average, robust SE	Robust SE
Wrongly punished$_{i,t-1}$	-0.643	-0.756	-0.625	-0.932	-1.049	-0.694	-0.582
	[0.257]*	[0.284]**	[0.271]*	[0.412]*	[0.565]	[0.412]	[0.247]*
False Negatives Treatment × *wrongly punished*$_{i,t-1}$							-0.206
							[0.468]
Wrongly unpunished$_{i,t-1}$	-0.486	-0.365	-0.548	-0.973	-0.839	-0.933	-0.610
	[0.234]*	[0.262]	[0.231]*	[0.270]**	[0.29]**	[0.269]**	[0.205]**
False Negatives Treatment × *Wrongly unpunished*$_{i,t-1}$							-0.236
							[0.285]
Contribute$_{i,t-1}$	1.173	0.475	0.814	0.890	0.000	0.667	0.992
	[0.207]**	[0.25]	[0.221]**	[0.220]**	[0.255]	[0.230]**	[0.146]**
Average total contributions in group$_{i,t-1}$			0.616			0.558	
			[0.144]**			[0.163]**	
One other contributor$_{i,t-1}$	0.774	0.703	0.514	-0.130	0.258	-0.148	0.497
	[0.248]**	[0.263]**	[0.256]*	[0.359]	[0.406]	[0.359]	[0.194]*
Two other contributors$_{i,t-1}$	1.161	0.976	0.574	0.326	0.779	0.031	0.939
	[0.253]**	[0.303]**	[0.289]*	[0.343]	[0.404]	[0.353]	[0.193]**
Three other contributors$_{i,t-1}$	1.867	1.528	0.815	0.743	1.237	0.214	1.434
	[0.298]**	[0.41]**	[0.391]*	[0.356]*	[0.431]**	[0.382]	[0.211]**
False Negatives Treatment							0.429
							[0.137]**
Observations	836	836	836	836	836	836	1672

Standard errors in brackets. * indicates significant at 5%; ** significant at 1%. Two-tailed tests.
Coefficients for constant and period indicators suppressed in all columns to save space.

Table 4: Punishment and Contribution Rates in Technology Choice False Positives Condition

Punishment Rate Given Accusatory Signal	Punishment Rate Given Exculpatory Signal	Difference in Punishment Rates	Contribution Rate
Technology Choice False Positives			
1.000	0.000	1.000	0.825
0.939	0.000	0.939	0.738
0.730	0.000	0.730	0.775
0.652	0.000	0.652	0.238
0.635	0.000	0.635	0.363
0.619	0.059	0.560	0.400
0.419	0.000	0.419	0.688
0.435	0.091	0.344	0.275
0.526	0.250	0.276	0.138
0.538	0.500	0.038	0.025

Notes: Each entry is a single group observed over twenty periods. Sorted by difference in punishment rates.

Table 5: Predicting the decision to contribute in False Positives Treatment and Technology Choice False Positives Condition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Technology Choice False Positives			Pooled, Technology Choice False Positives and False Positives			
	Robust SE	Random group and individual effects	With historical average, robust SE	Robust SE	Random group and individual effects	With historical average, robust SE	Random effects and historical average
Wrongly punished$_{i,t-1}$	-0.654	-0.856	-0.754	-0.760	-0.850	-0.725	-0.819
	[0.284]*	[0.321]**	[0.291]**	[0.253]**	[0.276]**	[0.270]**	[0.274]**
Tech. Choice False Positives				0.273	0.188	0.178	0.194
$\times$ *Wrongly punished*$_{i,t-1}$				[0.350]	[0.388]	[0.361]	[0.382]
Wrongly unpunished$_{i,t-1}$	-1.136	-0.926	-1.047	-0.332	-0.232	-0.408	-0.274
	[0.302]**	[0.337]**	[0.308]**	[0.217]	[0.241]	[0.218]	[0.239]
Tech. Choice False Positives				-0.855	-0.764	-0.699	-0.765
$\times$ *Wrongly unpunished*$_{i,t-1}$				[0.327]**	[0.372]*	[0.330]*	[0.365]*
Contribute$_{i,t-1}$	1.673	0.831	1.182	1.381	0.616	0.980	0.546
	[0.229]**	[0.275]**	[0.248]**	[0.150]**	[0.182]**	[0.162]**	[0.183]**
Average total contributions			0.911			0.699	0.570
in group$_{i,t-1}$			[0.175]**			[0.103]**	[0.157]**
One other contributor$_{i,t-1}$	1.375	0.894	0.702	1.015	0.744	0.613	0.703
	[0.276]**	[0.345]**	[0.307]*	[0.180]**	[0.204]**	[0.191]**	[0.203]**
Two other contributors$_{i,t-1}$	1.788	0.975	0.442	1.382	0.927	0.554	0.805
	[0.288]**	[0.431]*	[0.385]	[0.184]**	[0.239]**	[0.222]*	[0.244]**
Three other contributors$_{i,t-1}$	2.512	1.486	0.700	2.073	1.400	0.788	1.191
	[0.323]**	[0.516]**	[0.465]	[0.215]**	[0.306]**	[0.288]**	[0.318]
Tech. Choice False Positives				0.167	0.077	0.281	0.268
				[0.147]	[0.440]	[0.151]	[0.270]**
Observations	760	760	760	1596	1596	1596	1596

Standard errors in brackets. * indicates significant at 5%; ** significant at 1%. Two-tailed tests.
Coefficients for constant and 18 period indicators suppressed in all columns to save space.

Figure 1: Average Group Contribution Levels by Period, False Positives and False Negatives Treatments

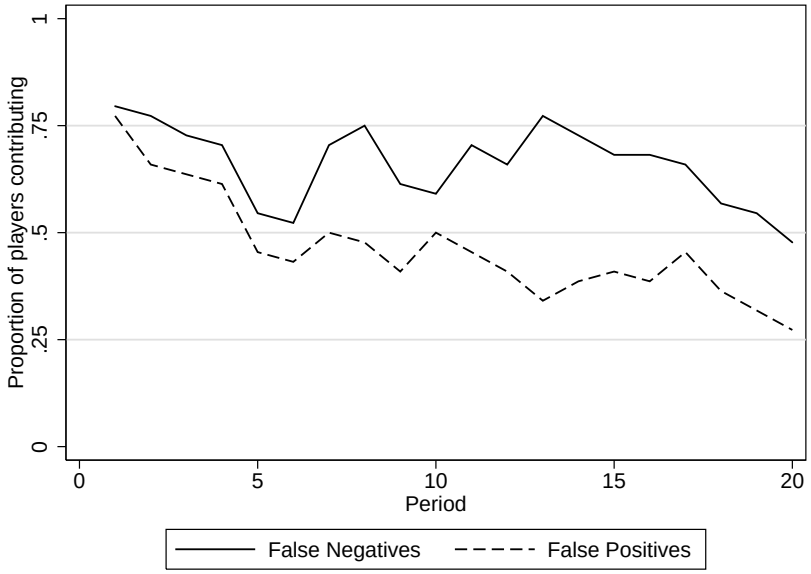
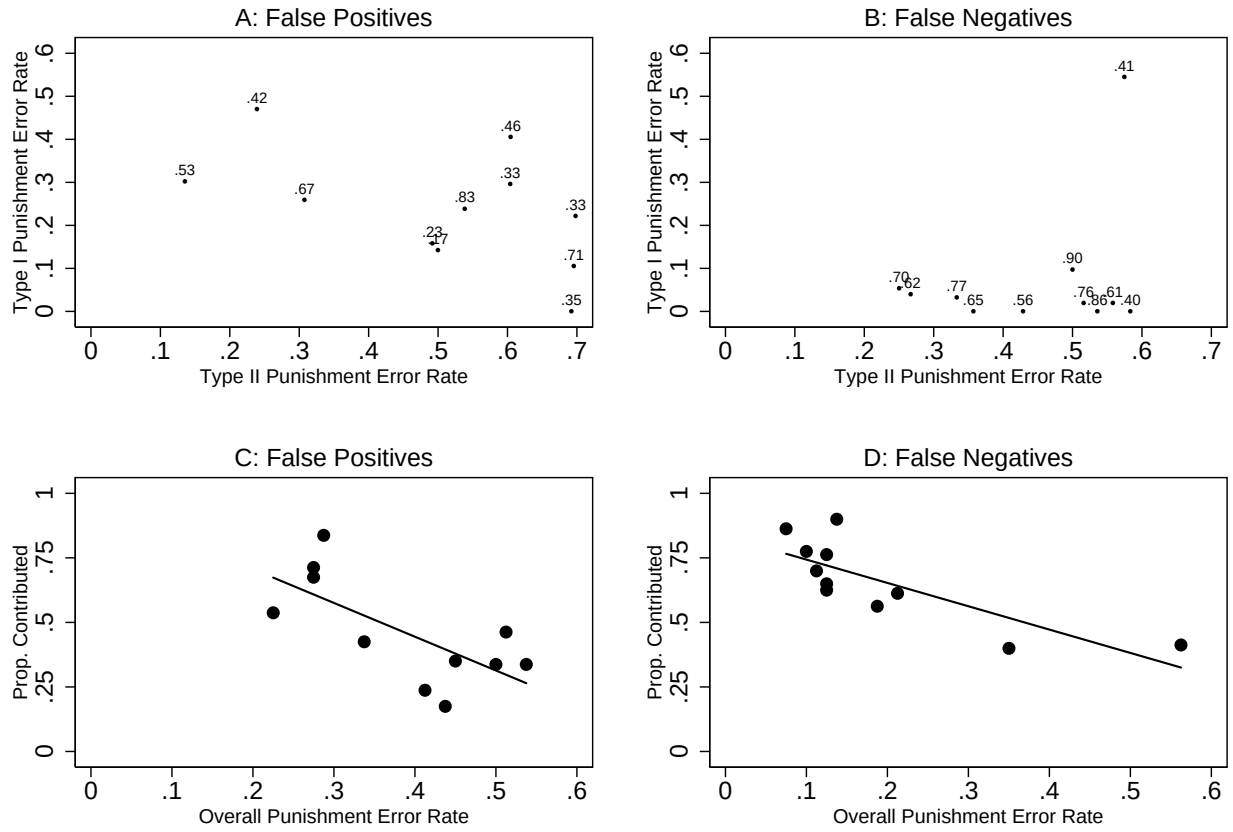
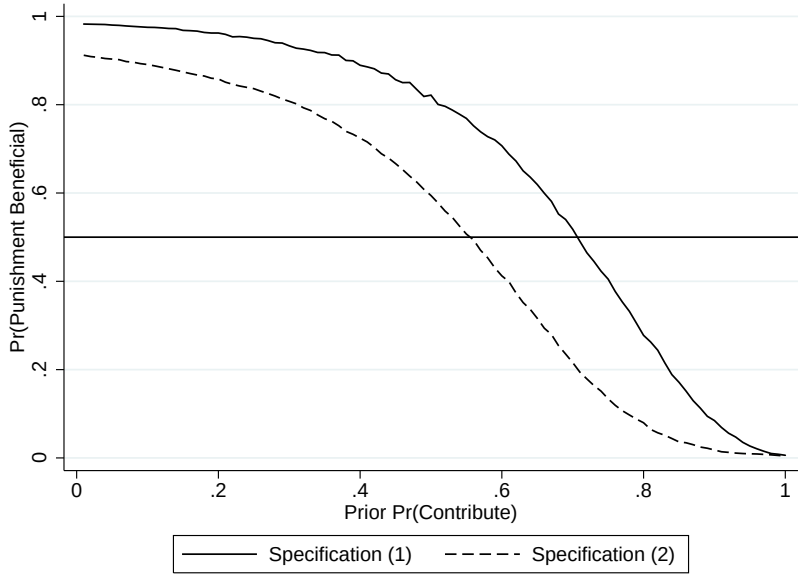


Figure 2: Average Contribution and Error Rates by Group, False Positives and False Negatives Treatments



Note: Numbers in panels A and B are average contribution rates in group. Lines in panels C and D are predicted values from linear regression.

Figure 3: Estimated Probability that Punishment Given Accusatory Signal Improves Subsequent Compliance, False Positives Treatment



Note: Specifications refer to simulations run using parameter estimates shown in those columns of Table 3.

Figure 4: Average Group Contribution Levels by Period, False Positives Treatment and Technology Choice False Positives Condition

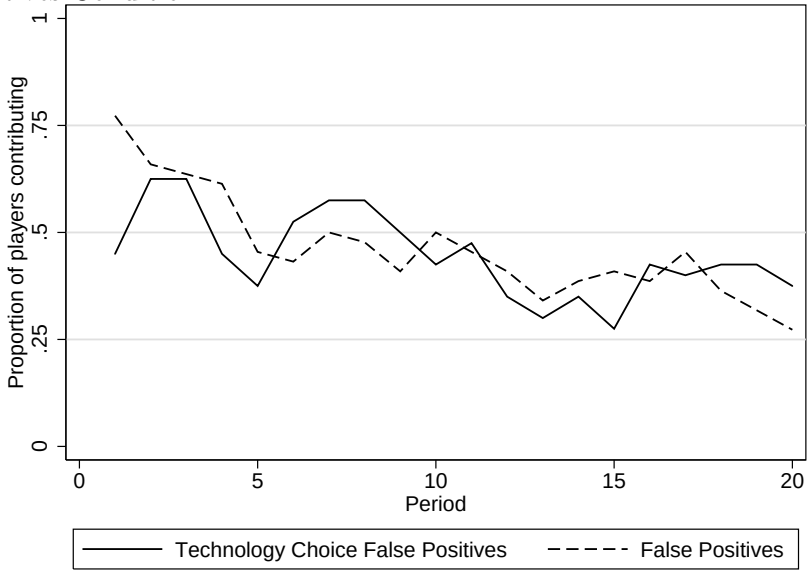
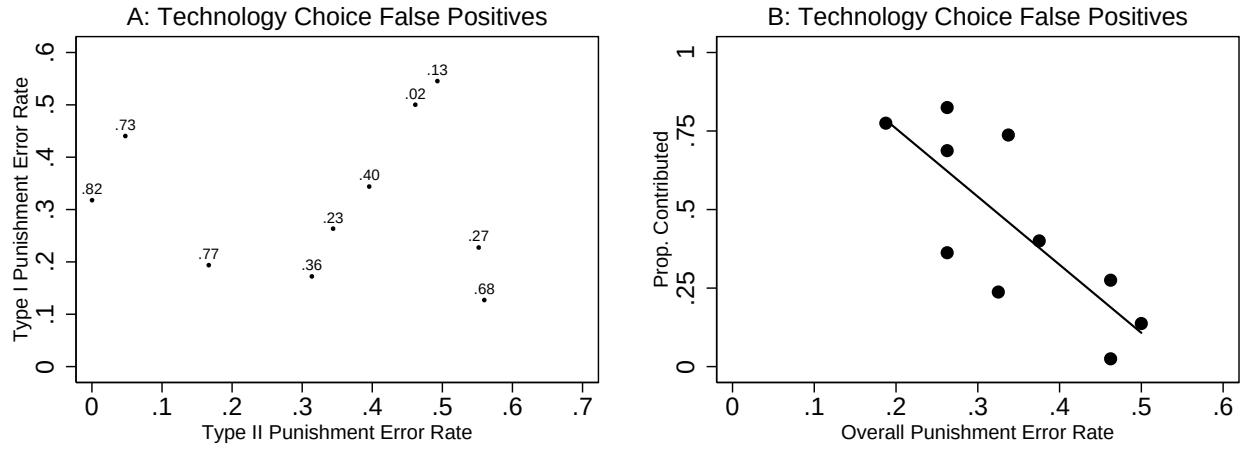
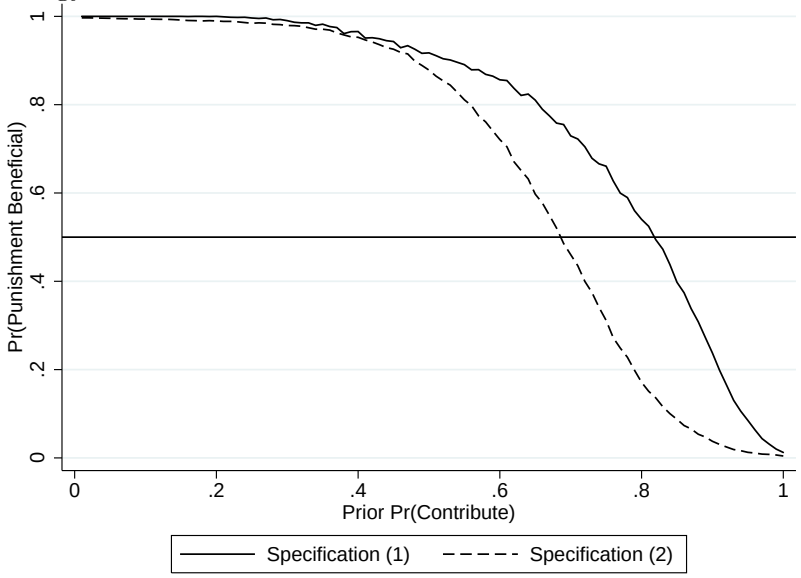


Figure 5: Average Contribution and Error Rates by Group, Technology Choice False Positives Condition



Note: Numbers in panel A are average contribution rates in group. Line in panel B is predicted values from a linear regression.

Figure 6: Estimated Probability that Punishment Given Accusatory Signal Improves Subsequent Compliance, Technology Choice False Positives Condition



Note: Specifications refer to simulations run using parameter estimates shown in those columns of Table 5.

Figure 7: Monitor Strategies and Average Contribution Rates, Across Three Treatments