Procedural Justice and Legitimate Authority*

June 26, 2014

Eric Dickson  Sanford C. Gordon  Gregory A. Huber
New York University  New York University  Yale University
eric.dickson@nyu.edu  sanford.gordon@nyu.edu  Gregory.huber@yale.edu

Abstract

Social psychologists (e.g., Tyler 1990) have argued that an authority’s use of fair procedures can improve compliance by enhancing the authority’s legitimacy. However, previous efforts to assess this effect empirically are confounded because procedures that enhance legitimacy also alter subjects’ perceptions of the authority’s efficacy in sanctioning noncompliance. We evaluate the procedural legitimacy argument in the context of two public goods experiments. In the first, following subjects’ contribution choices, an enforcer makes a costly procedural choice of “high accuracy information” that improves the likelihood that “the punishment fits the crime” should she endeavor to punish a non-contributor. Then, the game is repeated. By design, contributors and non-contributors alike will be exposed to the possibility of punishment under each procedure. This permits us to estimate a “specific legitimacy” effect of the authority’s procedural choice on the subsequent contribution behavior of a punished citizen. In the second experiment, the procedure is randomly assigned to, rather than chosen by, the authority. We find strong evidence of a specific legitimacy effect in the study with endogenous accuracy choice, and no corresponding effect with exogenous institutions. The experiment underscores the difficulty in estimating “general legitimacy” effects even in experimental settings, as well as the difficulty of comparing instrumental and non-instrumental motivations for compliance with prescribed behavioral norms.

* We thank Hannah K. Simpson and Sebastian Thieme for superlative research assistance.
Introduction

Two different police officers stop two different gang members because they match the description of the suspects in a nearby robbery. In one interaction, the police officer carefully follows department procedure and queries the youth extensively about where he was earlier that day. She finds his answers unpersuasive and inconsistent, and arrests him. In the other interaction, the officer is rude and quick to judge. She arrests the youth after deciding his resemblance to a witness’s description of a suspect is “close enough.” In the first interaction, the officer behaves in a manner consistent with norms of procedural fairness, whereas in the second the officer does not. If both suspects are actually guilty, does the fact that one officer follows better procedures affect the likelihood that the punished offender recidivates? And if it does, why?

In the several decades since the seminal study of Thibaut and Walker (1975), scholars have explored the relationship between citizens’ evaluations of an authority’s commitment to procedural justice and their attitudes toward, and willingness to comply with, that authority (e.g., Lind and Tyler 1988; Tyler 1990; Tyler and Lind 1992; Tyler and Huo 2002; Paternoster et. al. 1997; Murphy 2004; Colquitt 2001). The channel through which this relationship is most often hypothesized to operate is legitimacy: a “judgment by group members that they ought to voluntarily obey social rules and authorities irrespective of the likelihood of reward or punishment” (Tyler 1997; see also French and Raven 1959). According to this account, procedurally fair institutions or behavior enhance an authority’s legitimacy. Enhanced legitimacy, in turn, motivates citizens to comply.

This paper identifies and addresses a fundamental problem of interpretation in prior research claiming empirical support for this legitimacy-based account: procedures or choices that enhance an authority’s legitimacy may simultaneously alter citizens’ beliefs about the authority’s capacity to bestow rewards and punishments. In the example above, the first police officer takes pains to follow a procedure that, by generating more accurate information, is fairer than the one adopted by her counterpart (cf. Leventhal 1980). The first youth may be more likely to comply than the second because the relative fairness of the first officer makes him perceive her, or the criminal justice system more broadly, as legitimate.

However, the first officer’s actions, which reduce the risk of accidentally punishing the innocent or freeing
the guilty, also enhance deterrence: the first youth may be more likely than the second to conclude that committing another crime in the future will lead to punishment; moreover, the second might conclude that the police are just looking to arrest “people who look like me,” undermining the benefit of law-abidingness. Because these encounters teach the two youths different things about both the fairness and the effectiveness of policing, we cannot attribute differences in subsequent behavior to one or the other mechanism.

The experiments and data analysis we present in this paper overcome this interpretative ambiguity. Subjects assigned to the role of citizens choose to contribute or not to a common pot; each would benefit if everyone contributed, but has an incentive to free-ride on the others by withholding his or her own individual contributions. Then, a subject assigned to the role of centralized authority with more or less accurate information about the actual behavior of the citizens can choose to punish a citizen. The interaction is then repeated.

As we articulate below, this design allows us to isolate the extent to which changes in a subject’s compliance stemming from the more accurate institution emerge from enhancements to the authority’s legitimacy, net of instrumental motivations for compliance. In particular, by comparing citizens who experienced punishment to those who witnessed it under high and low accuracy institutions, we can recover the effect on subsequent compliance of what we call specific legitimacy: the deference granted to an authority for non-instrumental reasons by an individual who has directly experienced punishment. General legitimacy – the deference afforded to authorities by individuals who observe, but may not directly experience punishment – is not separately identified, although we can establish an upper bound on its effect among those who experience a common event (e.g. a guilty person being punished).

In our first experiment, the subject in the role of centralized authority chooses, at a cost to him or herself, whether or not to invest in the more accurate institution. Our second experiment is identical, except that accuracy is randomly assigned. Comparing across the results of the experiments allows us to distinguish the intrinsic effects of fairer institutions from the possibility that players react to the authority’s willingness to incur a cost to act more fairly (cf. Avolio and Locke 2002; De Cremer, Dijke, and Bos 2004; Knippenberg

1We follow many scholars in using public goods games as a metaphor for social norm compliance (Ledyard 1995)
This, in turn, permits us to further probe the psychological mechanism underlying the citizens’ reaction to procedural fairness as a source of enhanced legitimacy.

In addition to allowing us to isolate legitimacy-based motivations from instrumental concerns, features of our experiments also enable us to address other important criticisms of prior research on procedural fairness and legitimacy. For example, survey-based research (e.g., Tyler 1990) has documented a strong association between subjects’ perceptions of the fairness of the legal process and the intention to comply with the law in the future. As other scholars have noted (e.g., van den Bos 2001), however, perceptions of fairness may be correlated with other unobserved individual- and institutional-level factors that also explain compliance. Similarly, to the extent that these studies rely on self-reported compliance rather than the observation of actual behavior, they run the risk of measurement error. In prior work that experimentally manipulates fairness (e.g., by altering whether someone who is grading an exam reads one or more answers [e.g., Vermunt et. al. 1996]), the outcomes that are measured are affect toward an institution (is it fair? Is it just?), rather than whether the institution alters future behavior. By contrast, the design of our studies permits direct observation of compliance and noncompliance – contributions to the public good – that mimics real-world settings in which citizens must choose between intergroup cooperation and personal gain.

**Instrumental and Non-Instrumental Motivations for Compliance: Specific and General**

Our central research question is whether an increase in the procedural fairness of an institution increases future compliance by enhancing the institution’s legitimacy. While there are multiple ways in which an institution’s procedures may be fair (Leventhal 1980), in this paper we focus on its accuracy. In any hierarchical relationship in which an authority is endowed with the capacity to reward compliance and punish

---

2 Leventhal (1980) comprehensive set of criteria includes (1) Consistency (across persons in space and time); (2) Bias suppression (absence of a personal stake on the part of the authority adjudicating disputes); (3) Accuracy (in the gathering of information and its application to decisions reached); (4) Correctability (the opportunity to remedy previous mistakes, e.g., by appeal); (5) Representativeness (closely related to the notion of process control; that subjects believe that they have a say in the decision making apparatus); and (6) Ethicality (compatibility of the procedures with fundamental moral and ethical values). Research that experimentally manipulates dimensions apart from accuracy include Grossman and Baldassari 2012 and Dal Bo, Foster, and Putterman 2011.
noncompliance, greater accuracy may increase future compliance by increasing the legitimacy of the authority. Greater accuracy can also enhance deterrence by reducing the likelihood that an innocent person is punished or that a guilty person is not. For these reasons, even an experiment that documents a relationship between accuracy and compliance is not guaranteed to isolate the extent to which that relationship is mediated through changes in legitimacy (non-instrumental motivations) or deterrence (instrumental concerns). This inferential problem is displayed graphically in Figure 1. In the figure, solid boxes denote observables, and dashed boxes unobservables. Ignoring (for the moment) the box labeled “authority choice” and the arrows extending from it, the figure shows the two channels through which fair procedure can affect subsequent citizen behavior.

Figure 1 About Here

A first step in addressing this ambiguity is to be more precise about what we mean by legitimacy, which we do with reference to a distinction commonly drawn between two different kinds of deterrence: general and specific (Stafford and Warr 1993; Williams and Hawkins 1986). General deterrence is the apprehension of the severity or likelihood of punishment that motivates compliance with laws or social norms, across society (e.g., Beccaria 1764; Bentham 1830; Paternoster 1987). That is, general deterrence is shaped by those factors that alter my expectations about future punishment independent of the personal experience of having been punished. By contrast, specific deterrence refers to an additional motivation to comply associated with actually being punished—a subject’s own experience of punishment activates a specific sensitivity to the threat of punishment that might be absent for members of the broader population, who are only aware of the threat in an abstract sense (Nagin, Cullen, and Johnson 2009).

We adopt the language of general and specific to discussions of legitimacy. In particular, we define general legitimacy as the deference accorded to an authority across the population as a consequence of non-instrumental considerations such as its fairness. So, for example, law-abiding citizens in the community where the police officer carefully evaluates evidence may hold that officer in higher esteem than law-abiding citizens in the community patrolled by her quick-to-judge counterpart. This esteem may, in turn, translate into a greater willingness to comply or cooperate with the officer.
By contrast, *specific legitimacy* refers to the deference accorded to the authority by the individual experiencing punishment (or reward) firsthand on the basis of non-instrumental considerations such as fairness, *above and beyond* general legitimacy.

Returning to the example that opens the paper, if the first youth is more likely to comply with the law in the future than the second, to what extent is this driven by his perceptions of the legitimacy of legal authority brought about by his procedurally fair treatment by the first police officer? Similarly, suppose both youths were innocent and wrongly arrested. Do the fair procedures followed by the first police officer preserve the legitimacy of law enforcement in a way that mitigates the factual injustice of the arrest? The experiments and research design we describe below allow us to measure the effect on subsequent compliance of more accurate (i.e., in expectation less likely to cause enforcement mistakes, and therefore, more fair) procedures that operate through enhanced specific legitimacy channel.

**Do Institutions or Choices Create Legitimacy?**

An additional question concerning legitimate authority is whether it arises because of exogenous features of the institution or as a consequence of choices made by the authority herself. In one view, institutions that reduce the likelihood of errors should be seen as more legitimate, and this should directly improve compliance. The experimental studies most similar to ours manipulate accuracy in this way, by assigning subjects to a condition in which an authority evaluates how well or how poorly the subject performed on a questionnaire by grading one, several, or all items. (Vermunt et. al. 1996; van den Bos 2001; De Cremer 2004).

Another perspective, however, is that it may be the fact that an authority chooses a superior institution that imbues it with legitimacy – particularly if that choice is costly to the authority. Returning to Figure 1, we see that the authority’s choice can affect the procedure, which in turn affects citizen behavior – either through the instrumental or non-instrumental channel; alternatively, the choice itself affects behavior through one or the other channel.

Our first study departs from prior research by having the accuracy of an institution be a consequence of a costly choice by an authority that is also responsible for punishment. Thus, the design implicates the notion of leader “self-sacrifice” as a legitimating factor (De Cremer 2006). In our second study, by contrast,
the accuracy of the institution is randomly assigned. Comparing the estimates of specific legitimacy across the two studies therefore allows us to understand what proportion of the specific legitimacy effect is innate to an institution and how much is driven by the costly action of the authority.

**Study 1: Specific Legitimacy and Costly Leader Choice**

**Method**

*Participants and Design:* We conducted five experimental sessions at the experimental social science lab of a major American university. Each of the 80 subjects who participated took part in one session only. Subjects interacted anonymously via networked computers. Participants signed up via a web-based recruitment system that draws on a large, pre-existing pool of potential undergraduate subjects. (Subjects were not recruited from the authors’ courses and did not receive course credit for participating.) After giving informed consent, subjects received written instructions that were subsequently read aloud to promote understanding and induce common knowledge of the experimental scenario. No deception was employed. Before beginning the experiment, subjects took an on-screen quiz that both measured and promoted understanding of the instructions. Subjects earned tokens, convertible into dollars at the end of the experiment (60 tokens = US$1), in amounts determined by the outcomes of play. Subjects’ overall payoffs in a given session were equal to the sum of the payoffs from each of fifteen periods in which they interacted, plus a US$7 show-up fee. The average participant earned about $18 for a session (including the show-up fee) that took approximately 1.5 hours.

*Procedure:* The game is a variant of a public goods game with a centralized authority (Fehr and Fischbacher 2004; Bernhard, Fehr, and Fischbacher 2006; Dickson, Gordon, and Huber 2009, forthcoming). We follow many scholars in using public goods games as a metaphor for social norm compliance (Ledyard 1995). At the beginning of each of the fifteen periods, subjects were randomly assigned to a group of five people, of which four were randomly assigned as citizens (“Role A” as described in the instructions and on-
screen) and one as an authority (“Role B”). Roles were constant for the duration of the period. Each period consisted of four separate “stages”:

(1A) Each citizen was randomly assigned a temporary ID number, and made a binary choice to contribute (to a “common pot”) or keep his/her entire initial endowment of 20 tokens.

(1B) The authority chose whether to receive “low accuracy” information (at no cost) or “high accuracy” information (at a cost of two tokens) about citizen contribution behavior. The authority then received information (more below) and had the opportunity to reduce the payoffs of a selected citizen by 20 tokens. Citizens then learned of the authority’s accuracy choice and the information the authority received about the behavior of the punished citizen (if any citizen was punished). All players also learned the total number of citizens who contributed to the public good and the actual contribution behavior of any citizen who was punished.

(2A) Stage (1A) was repeated (with citizens receiving new temporary ID numbers).

(2B) Stage (1B) was repeated.

(Because of the repetition, we will refer below to decisions made at stage 1A and 2A as “first stage” choices, and decisions made at stage 2A or 2B as “second stage” choices.) Note that when making their first stage contribution decisions, citizens did not know whether or not the authority would invest in high quality information. Similarly, when choosing whether to invest in superior information, the authority did not know how many, or which, of the citizens had contributed.

If the authority chose “low accuracy,” the information she received for each citizen was correct 60% of the time, and incorrect 40% of the time. By contrast, if she chose “high accuracy,” the information she received was correct 80% of the time, and incorrect 20% of the time.

The “common pot” to which individual citizens could contribute their endowment took the form of a standard linear public good: each citizen received 0.4 times the total amount allocated by all citizens. So, for example, a citizen who kept his tokens when the other three citizens allocated to the common pot would receive, in that stage, $20 + 0.4 \times 60 = 44$ tokens. As is standard in such games, citizens would be better off if
all citizens contributed (which would yield a payoff of 32 tokens to each citizen), but each has an individual incentive to “free ride” by keeping his or her 20 tokens for himself.

The authority received a fixed endowment of 20 tokens each period, and also received 0.4 times the total amount allocated to the common pot by all citizens. Thus, the authority was incentivized to do what she could to improve contributions, which would also benefit the citizens, and this incentive was commonly-known to all participants (as was the two-token cost of high-accuracy information).

As a point of comparison, we briefly describe the game-theoretic benchmark (GTB) prediction, employing the solution concept of subgame perfect Nash equilibrium. In stage 2B, the authority is indifferent over targeting strategies, and may choose any. Since there is no prospective benefit to the authority of investing in high accuracy information, she will not do so. Anticipating this, citizens will not contribute in stage 2A. (Critically, this holds for any targeting strategy the authority might adopt.) The same predictions hold for stages 1B and 1A. A more formal treatment of the benchmark can be found in Appendix A.

To be sure, the GTB relies on assumptions about behavioral rationality and the absence of intrinsic or relational motivations that are implausible in the current context. If citizens are more willing to comply with an authority that they perceive as legitimate, and the authority can improve her legitimacy through her investment in high-accuracy information, then she may choose to do so, especially in the first stage. Alternatively, the authority may have intrinsic motivations to pursue justice, even at a personal cost; such authorities may choose to invest in high-accuracy information at both opportunities to do so.

Data Overview

Our empirical strategy, described in detail below, entails estimating quantities using restricted subsamples of our data on citizen behavior. The complete dataset consists of 960 observations, where the unit of analysis is the citizen-period. Table 1 presents a partition of these observations into first-stage histories relevant to our analysis. For example, Row (1) of the table encompasses citizens who contributed in the first stage and whose group saw correct punishment: that is, punishment of a non-contributor following a signal to the authority that the target had, in fact, not contributed. Within this row, citizens are further subdivided into those who directly experienced punishment or merely witnessed another group member being punished;
and into citizens whose authority chose the low accuracy or the high accuracy institution. Because a
contributor can’t herself be correctly punished, the “experienced” cells are blank. Likewise, a non-contributor
cannot experience incorrect punishment, so the corresponding cells in Row (4) are blank. For the rows
featuring incorrect punishment (i.e., punishment of a contributor), we include only cases in which the
authority punished the contributor after receiving a signal that the latter had not contributed. The handful of
cases in which the authority targeted “perversely” – punishing someone about whom she had received a
“contributed” signal – are excluded from Row (4) and appear in Row (6) of the Table.

Table 1 About Here

Several interesting features of the Table are worthy of note. First, authorities who invested in high
accuracy information in the first stage were far more willing to punish than those that did not. Second,
perverse punishment is rare, and particularly so following investment in high accuracy.

*Empirical Strategy and Measures*

In each of the analyses we conduct, we are interested in estimating the effect of the authority’s first
stage investment in high-accuracy information on subsequent second stage citizen contributions. As we noted
above, the investment can improve subsequent compliance by enhancing deterrence or legitimacy. Here, we
describe our strategy for distinguishing between these two causal pathways.

Our approach exploits a differences-in-differences design, whose basic logic is summarized in Table
2. We begin by focusing on cases in which a non-contributor in the first stage was correctly punished or
witnessed the correct punishment of another non-contributor. This is the subset of the data displayed in the
second row of Table 1. Within this sample, there are four kinds of citizens: those who directly experienced
punishment under the high accuracy institution (the top-left cell of Table 2); those who witnessed (but did
not experience) punishment under high accuracy (top-right); those who experienced punishment under low
accuracy (bottom-left); and those who witnessed punishment under low accuracy (bottom-right).

Table 2 About Here

Analytically, the factors that predict second-stage contribution behavior for those who experienced
punishment under high accuracy in the first stage (top-left) are (1) *intrinsic motivation* to contribute; (2) *group*
norms (more on this below); (3) specific deterrence associated with the experience of punishment; (4) beliefs about the likelihood of future punishment associated with the high-accuracy choice (general deterrence [high]); (5) The authority’s general legitimacy given the high accuracy choice (general legitimacy [high]); and (6) the specific legitimacy effect associated with being (appropriately) punished by an authority who invested in high accuracy (specific legitimacy [high]).

For those non-contributors who witnessed punishment of a non-contributor in the high accuracy environment (top-right), four of the above factors are operative: (1), (2), (4), and (5). Specific deterrence and specific legitimacy are excluded because these players did not directly experience punishment.\(^5\)

The same logic governs the operative factors for citizens who experienced and witnessed punishment in the first stage when the authority did not invest in more accurate information, replacing [high] with [low] to indicate the lower accuracy environment.

Table 1 puts the problem of empirically distinguishing instrumental and non-instrumental motivations for compliance in stark relief. Suppose we were to adopt a standard cross-sectional approach, comparing those individuals punished under high accuracy with those punished under low accuracy. This would entail restricting attention to the left column, and estimating the sample analog of

\[ D_{op} = EC_2(\text{high accuracy, experienced}) - EC_2(\text{low accuracy, experienced}), \]

where \( EC_2 \) refers to expected second-stage contribution. As is evident from the Table, this yields

\[ D_{op} = (\text{general legitimacy [high]} - \text{general legitimacy [low]}) + \]
\[ (\text{specific legitimacy [high]} - \text{specific legitimacy [low]}) + \]
\[ (\text{general deterrence [high]} - \text{general deterrence [low]}). \]

\(^5\) Suppose a specific legitimacy effect exists, that is that a player being correctly punished under high accuracy is subsequently more likely to contribute for non-instrumental reasons. We assume that the other players in the group do not anticipate this reaction, which if citizens base their contributions about the contributions of others, would manifest as a common increase in compliance for all citizens. If they did anticipate the reaction and were therefore also more likely to comply, then this would tend to deflate the specific legitimacy estimate and inflate the combined effects of general legitimacy and deterrence. As we show below, we find no evidence for a general legitimacy/deterrence effect for investment in high accuracy, reducing concerns about this spillover effect downwardly biasing the specific legitimacy estimate.
In words, the cross-sectional difference, \(D_{op}\), is a biased measure of the legitimating effect of the procedurally fair institution: it returns the effect of the institution on behavior stemming through the general and specific legitimacy channels, but also through the instrumental general deterrence channel.

Our approach instead leverages the distinction between specific and general legitimacy to isolate a component of the effect of the institution stemming exclusively from its legitimating effect. The differences-in-differences strategy we adopt proceeds by estimating the sample analog of

\[
DD = (EC_2(\text{high accuracy, experienced}) - EC_2(\text{high accuracy, witnessed})) - (EC_2(\text{low accuracy, experienced}) - EC_2(\text{low accuracy, witnessed})).
\]

As is evident from the Table, this yields specific legitimacy [high] – specific legitimacy [low], that is, the incremental effect on second stage contributions of the more accurate informational environment as operating through enhanced specific legitimacy.

This differences-in-differences approach can be embedded in a regression framework. Specifically, using the sample of citizens described in the first row of Table 1 (non-contributors who experienced or witnessed warranted punishment), we regress second-stage contribution choice on an accuracy indicator variable (1=high, 0=low), a punishment experience indicator (1=experienced, 0=witnessed), and the multiplicative interaction of the two. The coefficient on the interaction term is the estimate of \(DD\). The chief advantage of the regression approach is that it permits us to control for the first stage contribution level, which plays a critical role in establishing group norms for second-stage contributions. Lacking a clear \textit{a priori} expectation about functional form, we adopt the conservative approach of including indicator variables for levels of first stage contributions.

Before proceeding, we note two additional features of this design. First, it is not possible to distinguish a general legitimacy from a general deterrence effect. Thus, \(DD\) represents a lower bound on the total (specific + general) legitimacy effect given any enforcement outcome (e.g., a non-contributor is punished). Note, however, that the difference

\[
EC_2(\text{high accuracy, witnessed}) - EC_2(\text{low accuracy, witnessed})
\]
yields the sum of the incremental general deterrence and general legitimacy effects. (The regression analog is the coefficient on the accuracy indicator.)

Second, while the above description has focused on non-contributors witnessing or experiencing punishment, we can deploy the same strategy applies to cases in which a contributor is punished (wrongly). That is, we can restrict our attention to contributors and compare those who are punished to those who witnessed punishment. These are the observations from the third row of Table 1. This comparison allows us to recover the specific legitimacy effect, and the sum of general legitimacy and general deterrence, associated with incorrect punishment.

Results

Total Effect of Accuracy Investment. We first examine the effect of the authority’s first stage investment in high accuracy on the subsequent contribution choices of each citizen in an authority’s group. These effects capture the combined influence of instrumental and non-instrumental motivations. The leftmost pair of bars in Figure 1 display average second-stage contributions under low (pale gray) and high (dark gray) accuracy for all observations except for those following perverse punishment (see above), along with bootstrapped 95% confidence intervals. As is evident from the figure, the authority’s investment in high accuracy is associated with a highly statistically significant 13 percentage point increase in the probability a player contributed (from 25 to 37%). This represents a nearly 50% increase in the second-stage contribution rate overall. The second set of bars repeats the comparison, excluding instances in which no first-stage enforcement took place. The result is a still-highly significant 10.2 percentage point increase.

The next two pairs of columns restrict attention to groups in which either correct or incorrect punishment took place. Our main analysis below, which relies on this partition of the data, may be seen as a decomposition of the total effects displayed in the Figure. As is evident from the figure, the effect of accuracy investment conditional on correct punishment is a large, highly significant 14.5 percentage points. However,

---

6 Regression-adjusted estimates employing a battery of first-stage covariates return nearly identical results.
the effect of the accuracy investment is both substantively and statistically insignificant when punishment is incorrect.

*Decomposing the Effect of Accuracy Investment: Correct Punishment.* Using our differences-in-differences empirical strategy, we now focus on the specific legitimacy effect of the authority’s investment in high accuracy *conditional on correct punishment.* This entails running the interactive regression specification described above on the sample of noncontributors who either witnessed or experienced correct, non-perverse punishment (Row 2 in the data partition described in Table 1).

Model estimates from a series of OLS regressions appear in the first four columns of Table 3. (Summary statistics for all variables appear in the Appendix.) In all specifications, the dependent variable is whether the player contributed in the second stage (1=yes, 0=no). In addition to indicator variables for high versus low accuracy and experienced vs. witnessed punishment, all specifications include indicator variables for the total number of first stage contributions. As a robustness check, odd columns include period-specific indicators (1-15) as well. Standard errors are clustered at the subject-level.7

Table 3 About Here

The first thing to note in the table is that across the four specifications, the first-order accuracy investment and experience coefficients are small and statistically indistinguishable from zero. The accuracy investment term is of particular substantive interest to us, because (as derived above) it represents the sum of the general deterrence and general legitimacy effects of the authority’s investment (conditional on correct punishment). The null finding for this variable suggests that the sum is negligible. If we impose the additional assumption that neither general deterrence nor general legitimacy is negative, then we can go further: the estimate represents an upper bound on the general legitimacy effect of the authority’s investment among non-contributors. Under this interpretation, the null finding on the upper bound implies there is not general legitimacy effect for these subjects.

7 Clustering at different levels (e.g., the group-period) or not at all has no discernible effect on the substantive findings. We also ran specifications including subject-specific fixed effects, which further confirmed the results reported here.
In each regression specification, the coefficient on the interaction term represents our estimate of the specific legitimacy effect of the authority’s investment among non-contributors. Across specifications, this effect is large and highly significant. Using both the column (1) and (2) specifications, we estimate about a 17 percentage point increase in the probability of a second stage contribution for first-stage noncontributors that is directly attributable to the specific legitimacy channel. When one considers that the average second-stage contribution across all citizens who did not contribute in the first stage is just 11.5%, the estimated specific legitimacy effect is particularly noteworthy. The comparison to our null finding for the general legitimacy effect is also suggestive, as it implies that an appreciation of the procedural fairness of a particular investment is more important for the individual directly experiencing it than one who simply observes its use.

Columns (3) and (4) reestimate the models from (1) and (2) by reweighting the data, so that each group/period receives the same weight. For example, in a group with two non-contributors, each has a weight of .5, whereas in a group with 4, each has a weight of .25. This approach yields a slight increase in the specific legitimacy effect (to about 19 points).

Decomposing the Effect of Accuracy Investment: Incorrect Punishment. Our next step is to implement our empirical strategy while restricting attention to the sample of citizens who contributed in the first stage and either witnessed or experienced the punishment of a contributor (corresponding to Row 3 in Table 1). Columns (5) through (8) of Table 3 display regression results for this sample; the specifications are analogous to those employed in columns (1) through (4).

The first thing to note in the table is that the sample size is considerably smaller when we restrict attention to contributors who were either wrongfully punished or who experienced wrongful punishment. This is driven by two mutually reinforcing effects. First, even under low accuracy, wrongful punishment is rare – for it to occur, the authority must not only receive a “false positive” signal; she must also target the citizen for whom she received that signal. Second, authorities who fail to invest in high accuracy information are reluctant to punish at all. (Prior work documents a similar behavioral pattern, in which the increased chance of an error reduces the probability of undertaking enforcement, see Dickson, Gordon, and Huber 2009).
Next, we turn to the first-order effects of the authority’s first stage investment in high accuracy information and the citizen’s direct experience with punishment. As was the case in our examination of correct punishment, there is no discernible effect of the direct experience of being punished. In the current context, this implies that directly experiencing wrongful punishment does not itself dramatically reduce subsequent compliance.

However, in specifications (6) and (8), we observe a large and statistically significant effect of the accuracy investment. (The effects in columns (5) and (7) are large, but imprecisely estimated.) The accuracy investment variable, recall, represents the sum of the general legitimacy and general deterrence effect of the authority’s investment choice for these first-stage contributors. The 36 percentage point effect hints at the possible existence of a strong legitimizing effect for contributors, in stark contrast to the nonexistent effect for noncontributors. However, our research design does not permit us to definitively state whether this effect is driven by changing perceptions of the authority’s deterrent capacity, and indeed this result is consistent with an interpretation grounded purely in instrumental motivations.

Lastly, we turn to the specific legitimacy effect of the authority’s investment for the incorrectly punished. The estimates of this effect are negative, but not statistically significant in any specification (the smallest p-value, in the column (6) specification, is about 0.15). How should one interpret these negative estimates? One way is to note the relationship between the (imprecisely estimated) negative specific legitimacy effect and the positive, specification-sensitive first order accuracy effect. The sum of these coefficients is never distinguishable from zero; in the specifications with period-specific effects, the sum is substantively close to zero as well. The implication of this fact is that any gains from general legitimacy and general deterrence associated with the authority’s informational investment are dissipated by the deleterious effect of the experience of wrongful punishment. In that case, while it appears that choosing a superior institution may increase the compliance of those who witness incorrect punishment, for those who experience it their behavior is no different than if they had experienced incorrect punishment in a less fair institution. Indeed, across all specifications, the predicted second-stage contribution of someone who experiences incorrect
punishment under low accuracy is never statistically distinguishable from that of someone who experiences it under high accuracy.

**Study 2: Externally Imposed Institutions**

An important question in research on procedural fairness and leadership, one underscored by our finding of a strong specific legitimacy effect for an authority when punishing non-contributors in Study 1, is whether it is the authority’s costly investment in a procedurally fair institution, or the institution itself, that facilitates compliance for non-instrumental reasons. To answer this question, in Study 2 we replace the authority’s choice of a level of accuracy with one in which that accuracy level is imposed, at random.

**Method, Empirical Strategy, and Measures**

*Participants and Design:* We conducted four experimental sessions at the same social science lab used for Study 1. These four sessions involved 70 unique subjects. The protocol for the experiment in Study 2 was identical to the protocol for Study 1, with one exception. Whereas in Study 1, the authority had the opportunity to pay to improve the quality of the information she received on the contribution behavior of the citizens. In Study 2, the computer determined the choice of “high accuracy” or “low accuracy” information. In particular, in Stage 1B, the computer chose “high accuracy” with 50 percent probability, and “low accuracy” with 50 percent probability. Accuracy was persistent across stages. In particular, if the authority received high accuracy information in Stage 1B, she would continue to receive high accuracy information in Stage 2B, and likewise for low accuracy information. The empirical strategy and measures are identical to those for Study 1.

*Data Overview*

Restricting attention to subjects assigned to the role of citizen in a given period, the Study 2 dataset consists of 840 citizen-period observations. In Table 4, we partition the data in a manner identical to the partition in Table 1 for Study 1. Compared to Table 1, the most striking feature of Table 4 is the frequency with which authorities declined to punish. In Study 1, authorities who chose low accuracy frequently declined to punish.

---

8 Insert description of subject demographics.
while those who paid the cost for high accuracy information nearly always punished (94% of the time). When the accuracy is randomly assigned, enforcement becomes less frequent, particularly under high accuracy. Indeed, authorities who received a high accuracy signal for free in Study 2 were seven times more likely to “sit on their hands” and decline to punish any citizen than they were when they had voluntarily paid for it in Study 1.\textsuperscript{9}

As above, the most relevant subsamples of the data are the set of citizens who failed to contribute and either experienced or witnessed correct punishment (Row 2, $N=215$) and the set of citizens who contributed and either experienced or witnessed incorrect punishment (Row 3, $N=89$).

**Results**

**Total Effect of Randomly-Assigned Accuracy:** Figure 2 displays differences in mean second stage contribution rates as a function of the authority’s first stage accuracy investment, for different subsamples of the data. In stark contrast to the results conveyed in Figure 1, we find no evidence of any effect on subsequent contributions of accuracy when it is randomly assigned. Unlike in Study 1, we found a mild covariate imbalance when comparing first stage contributions under high and low accuracy ($p=0.12$). Therefore, as a robustness check, we ran a series of regressions of second stage contributions on an accuracy indicator, first-stage contribution-specific effects, and period-specific effects. We find a modest five percentage point effect of high-accuracy information on second-stage contributions in the full sample ($p=0.056$, two-tailed); however, this effect disappears once we restrict attention to those that experienced any enforcement, or (separately) correct or incorrect punishment. Even using the strongest evidence for an effect of the randomly assigned institution, we find that it is less as half as large as the comparable effect (12 percentage points) when the institution is chosen by the authority.

\textsuperscript{9} Importantly, this cannot be explained with reference to differential learning about first stage contribution rates over the course of the fifteen rounds of each experiment: The average first stage contribution rate in Study 1 was 0.39 and it is a nearly identical 0.41 in Study 2.
Decomposing the Effect of Randomly-Assigned Accuracy: Correct Punishment. The main results for Study 2 appear in Table 5. Columns (1) through (4) implement our empirical strategy for citizens who either experienced or witnessed correct punishment (as presented in Row 2 of Table 4). As in Study 1, we observe negligible effects of both the first-order accuracy term, as well as the first-order experienced punishment term. The substantive interpretation of the former is that there is no apparent general legitimacy effect among noncontributors.

Table 5 About Here

Turning to the interaction term, which captures the specific legitimacy effect of the high accuracy institution, we see results very different from those in Study 1. In particular, for no specification can we reject the null hypothesis that there is no specific legitimacy effect. Unweighted estimates are very close to zero. The estimates from the specifications that weight each group equally are negative, but imprecisely estimated and close to zero. The conjunction of these findings and those from Study 1 suggest that to the extent that accurate procedures enhance an authority’s legitimacy, the mechanism through which this occurs is the authority’s costly investment in the procedure, and not the procedure itself.

Decomposing the Effect of Randomly-Assigned Accuracy: Incorrect Punishment. In columns (5) through (8) of Table 5, we restrict attention to the set of citizens who contributed and either observed or experienced incorrect punishment. Recall that for the analogous sample in Study 1, the authority’s choice of the high accuracy procedure did not lead directly to increased second stage contributions. This null result is replicated in the environment with assigned accuracy: in no specification is the first-order term on the accuracy variable statistically distinguishable from zero. As with Study 1, our analysis here is potentially underpowered (N=72 observations), which may explain why the magnitude and sign of the estimates are sensitive to the inclusion of period indicators.

By contrast, we observe a large, statistically significant effect on the first-order incorrect punishment variable. This estimate implies a very large negative effect of experiencing wrongful punishment (relative to simply witnessing it) when the authority is assigned low accuracy information.
Finally, we examine our measure of the specific legitimacy effect given incorrect punishment: the interaction of punishment experience and assigned accuracy. In no specification can we reject the null hypothesis of no specific legitimacy effect for this sample. As an auxiliary exercise, we also examine the sum of the interaction effect and the first-order incorrect punishment term. This represents the effect of incorrect punishment on subsequent contributions under high accuracy information. These sums are still large in magnitude, but statistically indistinguishable from zero at conventional levels. The implication is that the deleterious effect of incorrect punishment is diminished somewhat under high accuracy, although the point estimates are not distinguishable from those under low accuracy.

Overall, the results of Study 2 are consistent with the notion that “leader self-sacrifice” is a source of positive affect towards an institution. There is no increase in specific legitimacy associated with enforcement under high accuracy information when the accuracy is randomly assigned. In other words, it is the costly action on the part of the authority that generates the increase in specific legitimacy. By contrast, across studies, the level of accuracy appears to have no discernible direct effect on the efficacy of punishing those who comply.

Discussion

Some Potential Objections Addressed

We consider two potential objections to our empirical strategy. First, one might argue that our approach assumes, in part, that the factors that contribute to each individual’s decision to contribute in the second stage— intrinsic motivation, group norms, general deterrence, general legitimacy, specific deterrence, and specific legitimacy—are substitutes for one another, when in reality they might interact in ways that we do not account for. For example, it could be the case that subjects who initially have the greatest intrinsic motivation to contribute respond more to the legitimating effects of the authority’s accuracy choice.

We note that because we recover separate differences-in-differences estimates for discrete subsets of citizens (e.g., noncontributors who experience or witness correct punishment in the first stage; contributors who experience or witness incorrect punishment), our approach explicitly allows for these kind of interactions. Our approach also accounts for important behavioral differences that are indicators of type (as
captured by first stage contribution behavior) as well as factors that may affect subsequent cooperation, like the first round behavior of other players. Additionally, in a separate analysis (available from the authors on request), we ran our differences-in-differences regressions with subject-specific fixed effects, exploiting the fact that subjects played in multiple rounds. Including these effects controls for all time-invariant subject characteristics, including intrinsic motivation to comply. Our results including fixed effects are nearly identical to those reported above.

A second potential objection to our differences-in-differences design is that players who witness, rather than experience, punishment may respond to that event depending on how they think the punished player will respond. How might this affect our results? Suppose that a citizen who observes punishment anticipates that someone punished correctly under low information will be less likely to contribute in the second stage than someone punished correctly under high information. If the citizen who observed punishment under high information herself contributes more in anticipation of this effect, it will tend to bias against detecting a specific legitimacy effect in our design, because it will drive up the difference in contributions of the unpunished citizens in the high accuracy condition (relative to the low accuracy condition), making it less likely that the punished player's greater likelihood to contribute will be detectable in our statistical estimates. Thus, the presence of this bias would imply that our estimates of the specific legitimacy effect are smaller than the true effect.

Methodological Implications

This paper makes several methodological contributions to the study of legitimate authority and its relationship with procedural fairness. First, we demonstrate how difficult it is to isolate the effects of an authority’s legitimacy on subjects’ subsequent behavior relative to other instrumental concerns. We note that even with the benefit of a highly flexible experimental environment, we are not able to isolate the effect of general legitimacy from general deterrence. Thus, our experiment permits us to isolate a particular kind of legitimacy: what we refer to as specific legitimacy, rather than its more general counterpart.
To underscore the point, consider one of the potentially artificial features of the experiments described above: that even though the authority is uncertain about the “guilt” or “innocence” of the punishment target at the time of the punishment choice, the target’s fellow citizens know with certainty. As we show above, this feature of the experiment is critical because it allows us to control for the beliefs of the subjects that pertain to their instrumental motivations. Our empirical strategy therefore requires that both the target and the fellow citizens have access to precisely the same information about the competence, resoluteness, and efficiency of the authority to ensure their beliefs are equivalent. In observational studies, this feature of the environment is unlikely to hold.

Additionally, our study departs from the corpus of social psychological research on procedural fairness in its use of an incentivized laboratory environment in which subjects receive financial compensation for their performance in a game. The value of this setup is twofold. First, it permits us to establish a benchmark of purely instrumentally rational behavior against which to compare actual behavior. Because in the current setup, the citizen’s optimal behavior from a selfish perspective is never to contribute irrespective of the procedures and choices of the authority, we can more definitively attribute observed differences in contribution behavior to specific psychological motivations. Second, because subjects benefit or suffer materially from their actions and those of other subjects, our experimental environment approximates the sorts of compliance choices that individuals must make in their day-to-day lives.

Substantive Implications

Our novel empirical approach also allows us to make progress in understanding the psychological origins of institutional legitimacy. Specifically, we show that procedurally fairer institutions, operationalized here as those that are more accurate, enhance an authority’s legitimacy. Enhanced legitimacy, in turn, increases the efficacy of punishing citizens who fail to comply, and those effects are beyond those that arise due to changes in instrumental motivations. Enhanced legitimacy therefore improves the efficacy of enforcing the law against those who are its deserving target. But procedures alone do not intrinsically enhance legitimacy. Rather, it appears that a leader’s choice of a fairer institution at a cost to herself enhances her legitimacy and therefore
the efficacy of her actions. A necessary step for future work is to undertake additional research using different designs to understand whether fairer institutions also generate concomitant improvements in compliance via a general legitimacy mechanism, although the methodological hurdles we identify here imply these efforts may be difficult.

Our approach also highlights the value of differentiating between different varieties of legitimate authority. In the current context, we have articulated a distinction between specific and general legitimacy that parallels the distinction between specific and general deterrence. In one sense, this distinction is born of methodological necessity, since, as we describe above, we are unable to isolate a generalized legitimacy effect. In a more important sense, however, the distinction is completely consistent with our understanding of citizen obligations toward figures of authority: that a citizen who experiences fair treatment at the hands of a sanctioning authority may be feel a heightened sense of obligation toward that authority relative to his fellow citizens, even if both they and he are fully cognizant that the authority acted fairly.

Further, as has been noted in previous research, there are numerous dimensions of procedural fairness aside from accuracy. For example, the authority may or may not take actions that appear biased against a member of an out-group, or may or may not provide a procedural remedy for incorrect decisions. While the experiments described here focus on accuracy, their design is sufficiently flexible to accommodate any number of other conceptions of procedural fairness. A task for future research would to compare the legitimating effects of these different notions of fairness, and to see whether they are driven by the presence of the fairer procedure or the authority’s costly investment in it.

Finally, we note that while this particular experimental design cannot recover an estimate of the effect of generalized legitimacy, this does not mean that no such design is possible. Prior observational research is similarly limited, as we show that estimates of general legitimacy and general deterrence will tend to be confounded. A key task for researchers is therefore to identify novel designs that allow the analyst to distinguish general legitimacy and deterrence. These concerns aside, our research provides new evidence to support the theoretical supposition that legitimacy is a psychologically relevant predictor of individual compliance with an authority. In contexts in which authority figures make choices that affect the way citizens
evaluate their legitimacy, we show that punishment of the guilty is more efficacious when it is undertaken by an authority who has taken pains to enhance that legitimacy. Thus, in addition to any improvements in material motivations for compliance (e.g., deterrence), more legitimate institutions directly enhance the likelihood that those who are correctly punished for their wrongdoings comply in the future.
References


Bentham, J. (1830). *The Rationale of Punishment*.


Figure 1. Graphical Display of How Fair Procedures and an Authority’s Choices Can Affect Citizen Behavior
Figure 2. Total Effect of First-Stage Accuracy Choice on Second-Stage Contributions: Study 1 (Choice of Accuracy)

Note: All estimates exclude instances of perverse enforcement (see text for definition)
Figure 3. Total Effect of First-Stage Accuracy Choice on Second-Stage Contributions: Study 2 (Randomly Assigned Accuracy)

Note: All figures exclude instances of perverse enforcement (see text for definition)
Table 1. Partition of Sample of Citizen-Period Observations in the Accuracy Choice Experiment, Study 1

<table>
<thead>
<tr>
<th>First stage</th>
<th>Low Accuracy</th>
<th>High Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experience</td>
<td>Witness</td>
</tr>
<tr>
<td>(1) Correct Punishment (Contributed)</td>
<td>—</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>10.70%</td>
<td></td>
</tr>
<tr>
<td>(2) Correct Punishment (Didn't Contribute)</td>
<td>53</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>9.60%</td>
<td>18.10%</td>
</tr>
<tr>
<td>(3) Incorrect Punishment (Contributed)</td>
<td>21</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>3.80%</td>
<td>4.90%</td>
</tr>
<tr>
<td>(4) Incorrect Punishment (Didn't Contribute)</td>
<td>—</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>6.50%</td>
<td></td>
</tr>
<tr>
<td>(5) No Punishment</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td></td>
<td>39.90%</td>
<td></td>
</tr>
<tr>
<td>(6) Perverse Punishment</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>1.60%</td>
<td>4.90%</td>
</tr>
<tr>
<td>Total Cases</td>
<td>N=552</td>
<td></td>
</tr>
</tbody>
</table>

Note: Cell entries are number of observations and percentage of cases within accuracy condition.
Table 2: Factors Predicting Stage 2 Contributions and Difference in Differences Design for Recovering Specific Legitimacy

<table>
<thead>
<tr>
<th>Experienced</th>
<th>Witnessed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High Accuracy</strong></td>
<td></td>
</tr>
<tr>
<td>intrinsic motivation +</td>
<td>intrinsic motivation +</td>
</tr>
<tr>
<td>group norms +</td>
<td>group norms +</td>
</tr>
<tr>
<td>specific deterrence +</td>
<td>general deterrence [high] + general legitimacy [high]</td>
</tr>
<tr>
<td>general deterrence [high] + general legitimacy [high]</td>
<td></td>
</tr>
<tr>
<td><strong>Low Accuracy</strong></td>
<td></td>
</tr>
<tr>
<td>intrinsic motivation +</td>
<td>intrinsic motivation +</td>
</tr>
<tr>
<td>group norms +</td>
<td>group norms +</td>
</tr>
<tr>
<td>specific deterrence +</td>
<td>general deterrence [low] + general legitimacy [low]</td>
</tr>
<tr>
<td>general deterrence [low] + general legitimacy [low] + specific legitimacy [low]</td>
<td></td>
</tr>
</tbody>
</table>
# Table 3: Estimates of Specific Legitimacy, Study 1

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Invested in High Accuracy in Stage 1</strong></td>
<td>0.007</td>
<td>-0.008</td>
<td>0.024</td>
<td>0.015</td>
<td>0.150</td>
<td>0.369</td>
<td>0.148</td>
<td>0.360</td>
</tr>
<tr>
<td></td>
<td>[0.035]</td>
<td>[0.036]</td>
<td>[0.043]</td>
<td>[0.045]</td>
<td>[0.142]</td>
<td>[0.163]**</td>
<td>[0.148]</td>
<td>[0.162]**</td>
</tr>
<tr>
<td><strong>Punished (correctly) in Stage 1</strong></td>
<td>-0.052</td>
<td>-0.053</td>
<td>-0.076</td>
<td>-0.076</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.044]</td>
<td>[0.043]</td>
<td>[0.050]</td>
<td>[0.050]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Punished (correctly) in Stage 1 x High Accuracy</strong></td>
<td>0.174</td>
<td>0.175</td>
<td>0.189</td>
<td>0.189</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.064]**</td>
<td>[0.062]**</td>
<td>[0.071]**</td>
<td>[0.068]**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Punished (incorrectly) in Stage 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.047</td>
<td>0.056</td>
<td>0.015</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[0.134]</td>
<td>[0.131]</td>
<td>[0.151]</td>
<td>[0.151]</td>
</tr>
<tr>
<td><strong>Punished (incorrectly) in Stage 1 x High Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.309</td>
<td>-0.336</td>
<td>-0.184</td>
<td>-0.263</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[0.233]</td>
<td>[0.227]</td>
<td>[0.245]</td>
<td>[0.258]</td>
</tr>
<tr>
<td><strong>0 Contributions in group in Stage 1</strong></td>
<td>0.041</td>
<td>0.132</td>
<td>0.036</td>
<td>0.075</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.034]</td>
<td>[0.110]</td>
<td>[0.036]</td>
<td>[0.142]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1 Contributions in group in Stage 1</strong></td>
<td>0.101</td>
<td>0.196</td>
<td>0.095</td>
<td>0.139</td>
<td>0.283</td>
<td>0.497</td>
<td>0.253</td>
<td>0.504</td>
</tr>
<tr>
<td></td>
<td>[0.038]**</td>
<td>[0.116]*</td>
<td>[0.040]**</td>
<td>[0.148]</td>
<td>[0.287]</td>
<td>[0.291]*</td>
<td>[0.268]</td>
<td>[0.288]*</td>
</tr>
<tr>
<td><strong>2 Contributions in group in Stage 1</strong></td>
<td>0.149</td>
<td>0.215</td>
<td>0.144</td>
<td>0.157</td>
<td>0.290</td>
<td>0.420</td>
<td>0.291</td>
<td>0.427</td>
</tr>
<tr>
<td></td>
<td>[0.056]**</td>
<td>[0.120]*</td>
<td>[0.055]**</td>
<td>[0.154]</td>
<td>[0.106]**</td>
<td>[0.174]**</td>
<td>[0.111]**</td>
<td>[0.185]**</td>
</tr>
<tr>
<td><strong>3 Contributions in group in Stage 1</strong></td>
<td>0.280</td>
<td>0.356</td>
<td>0.281</td>
<td>0.309</td>
<td>0.732</td>
<td>1.060</td>
<td>0.727</td>
<td>1.083</td>
</tr>
<tr>
<td></td>
<td>[0.130]**</td>
<td>[0.162]**</td>
<td>[0.128]**</td>
<td>[0.172]**</td>
<td>[0.095]**</td>
<td>[0.146]**</td>
<td>[0.099]**</td>
<td>[0.167]**</td>
</tr>
<tr>
<td><strong>4 Contributions in group in Stage 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.577</td>
<td>0.721</td>
<td>0.570</td>
<td>0.757</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[0.212]**</td>
<td>[0.172]**</td>
<td>[0.203]**</td>
<td>[0.193]**</td>
</tr>
</tbody>
</table>

| Observations | 374       | 374       | 374       | 374       | 74        | 74        | 74        | 74        |
| R-squared    | 0.180     | 0.210     | 0.230     | 0.250     | 0.630     | 0.740     | 0.570     | 0.690     |
| Includes Period Indicators? | No      | Yes       | No        | Yes       | No        | Yes       | No        | Yes       |

Robust standard errors in brackets  
* significant at 10%; ** significant at 5%; *** significant at 1%
Table 4. Partition of Sample of Citizen-Period Observations in the Assigned Accuracy Experiment, Study 2

<table>
<thead>
<tr>
<th>First stage</th>
<th>Low Accuracy</th>
<th>High Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experience</td>
<td>Witness</td>
</tr>
<tr>
<td>(1) Correct Punishment (Contributed)</td>
<td>—</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.10%</td>
</tr>
<tr>
<td>(2) Correct Punishment (Didn't Contribute)</td>
<td>30</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>7.40%</td>
<td>12.10%</td>
</tr>
<tr>
<td>(3) Incorrect Punishment (Contributed)</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>5.00%</td>
<td>5.70%</td>
</tr>
<tr>
<td>(4) Incorrect Punishment (Didn't Contribute)</td>
<td>—</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.20%</td>
</tr>
<tr>
<td>(5) No Punishment</td>
<td>188</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>46.50%</td>
<td>41.30%</td>
</tr>
<tr>
<td>(6) Perverse Punishment</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>1.00%</td>
<td>3.00%</td>
</tr>
</tbody>
</table>

Total Cases

N=404
N=436

Note: Cell entries are number of observations and percentage of cases within accuracy condition.
Table 5: Estimates of Specific Legitimacy, Study 2

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Among non-contributors experiencing or witnessing correct punishment</td>
<td>Weighting Each Group/Period Equally</td>
<td>Among contributors experiencing or witnessing incorrect punishment</td>
<td>Weighting Each Group/Period Equally</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assigned High Accuracy in Stage 1</td>
<td>0.014</td>
<td>0.050</td>
<td>0.011</td>
<td>0.045</td>
<td>0.148</td>
<td>-0.142</td>
<td>0.110</td>
<td>-0.180</td>
</tr>
<tr>
<td>[0.059]</td>
<td>[0.065]</td>
<td>[0.063]</td>
<td>[0.068]</td>
<td>[0.109]</td>
<td>[0.151]</td>
<td>[0.111]</td>
<td>[0.160]</td>
<td></td>
</tr>
<tr>
<td>Punished (correctly) in Stage 1</td>
<td>-0.016</td>
<td>-0.027</td>
<td>0.040</td>
<td>0.010</td>
<td>-0.016</td>
<td>-0.027</td>
<td>0.040</td>
<td>0.010</td>
</tr>
<tr>
<td>[0.068]</td>
<td>[0.074]</td>
<td>[0.078]</td>
<td>[0.082]</td>
<td>[0.068]</td>
<td>[0.074]</td>
<td>[0.078]</td>
<td>[0.082]</td>
<td></td>
</tr>
<tr>
<td>Punished (correctly) in Stage 1 x High Accuracy</td>
<td>-0.006</td>
<td>0.012</td>
<td>-0.094</td>
<td>-0.045</td>
<td>-0.006</td>
<td>0.012</td>
<td>-0.094</td>
<td>-0.045</td>
</tr>
<tr>
<td>[0.079]</td>
<td>[0.084]</td>
<td>[0.091]</td>
<td>[0.092]</td>
<td>[0.109]</td>
<td>[0.151]</td>
<td>[0.111]</td>
<td>[0.160]</td>
<td></td>
</tr>
<tr>
<td>Punished (incorrectly) in Stage 1</td>
<td>-0.369</td>
<td>-0.415</td>
<td>-0.392</td>
<td>-0.468</td>
<td>0.079</td>
<td>0.195</td>
<td>0.070</td>
<td>0.263</td>
</tr>
<tr>
<td>[0.140]</td>
<td>[0.131]</td>
<td>[0.142]</td>
<td>[0.129]</td>
<td>[0.140]**</td>
<td>[0.131]**</td>
<td>[0.142]**</td>
<td>[0.129]**</td>
<td></td>
</tr>
<tr>
<td>Punished (incorrectly) in Stage 1 x High Accuracy</td>
<td>0.017</td>
<td>0.272</td>
<td>0.022</td>
<td>0.312</td>
<td>0.017</td>
<td>0.272</td>
<td>0.022</td>
<td>0.312</td>
</tr>
<tr>
<td>[0.048]</td>
<td>[0.251]</td>
<td>[0.051]</td>
<td>[0.211]</td>
<td>[0.196]</td>
<td>[0.236]</td>
<td>[0.224]</td>
<td>[0.251]</td>
<td></td>
</tr>
<tr>
<td>1 Contributions in group in Stage 1</td>
<td>0.153</td>
<td>0.395</td>
<td>0.155</td>
<td>0.431</td>
<td>0.153</td>
<td>0.395</td>
<td>0.155</td>
<td>0.431</td>
</tr>
<tr>
<td>[0.051]**</td>
<td>[0.247]</td>
<td>[0.054]**</td>
<td>[0.207]**</td>
<td>[0.283]**</td>
<td>[0.263]**</td>
<td>[0.281]**</td>
<td>[0.243]**</td>
<td></td>
</tr>
<tr>
<td>2 Contributions in group in Stage 1</td>
<td>0.120</td>
<td>0.355</td>
<td>0.117</td>
<td>0.388</td>
<td>0.120</td>
<td>0.355</td>
<td>0.117</td>
<td>0.388</td>
</tr>
<tr>
<td>[0.071]**</td>
<td>[0.270]</td>
<td>[0.073]</td>
<td>[0.231]*</td>
<td>[0.120]**</td>
<td>[0.176]**</td>
<td>[0.121]**</td>
<td>[0.194]**</td>
<td></td>
</tr>
<tr>
<td>3 Contributions in group in Stage 1</td>
<td>0.679</td>
<td>0.869</td>
<td>0.668</td>
<td>0.889</td>
<td>0.679</td>
<td>0.869</td>
<td>0.668</td>
<td>0.889</td>
</tr>
<tr>
<td>[0.169]**</td>
<td>[0.216]**</td>
<td>[0.156]**</td>
<td>[0.170]**</td>
<td>[0.120]**</td>
<td>[0.166]**</td>
<td>[0.116]**</td>
<td>[0.173]**</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>213</td>
<td>213</td>
<td>213</td>
<td>213</td>
<td>213</td>
<td>213</td>
<td>213</td>
<td>213</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.220</td>
<td>0.270</td>
<td>0.310</td>
<td>0.370</td>
<td>0.640</td>
<td>0.710</td>
<td>0.630</td>
<td>0.710</td>
</tr>
<tr>
<td>Includes Period Indicators?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Standard errors clustered at the subject level in brackets

* significant at 10%; ** significant at 5%; *** significant at 1%
Appendix A. Formal Analysis of Game-Theoretic Benchmark

June 26, 2014

For a description of the sequence of the game and the corresponding payoffs, see the main text.

Let $N$ denote the number of citizens, indexed by $i = 1, \ldots, N$; and $A$ denote the authority. Let $c^s_i \in \{0, 1\}$ denote citizen $i$’s choice at stage $s \in \{1, 2\}$ whether or not to contribute to the common pot (0=not contribute; 1=contribute). Let $a^s \in \{0, 1\}$ be the authority’s choice in stage $s$ of whether to receive low accuracy information ($a^s = 0$) or high accuracy information ($a^s = 1$). (In Study 2, $a^s$ is not a choice, but randomly assigned.) Let $\sigma^s_i \in \{0, 1\}$ denote the signal $A$ receives about $i$’s contribution choice in stage $s$, where $\sigma^s_i = 0$ is an “accusatory” signal that the citizen did not contribute, and $\sigma^s_i = 1$ is an “exculpatory” signal that the citizen did contribute. Let $p^s(a)$ denote the accuracy of the signals as a function of the authority’s investment choice, and let $r$ be the marginal per capita return on the public good. Let $y$ denote the size of the citizen’s endowment in each period, and $f$ the size of the penalty the authority can impose. Finally, let $t^s \in \{1, 2, \ldots, N, \emptyset\}$ denote the identity of the enforcement target ($\emptyset$ if the authority declines to sanction anyone).

In both experiments described in the paper, $N = 4$, $p^s(1) = 0.8$; $p^s(0) = 0.6$; $r = 0.4$, and $y = f = 20$.

We restrict attention to symmetric equilibria in which (a) all citizens play the same contribution strategy and in which (b) if the authority receives the same signal about two or more citizens, she targets them with equal probability. Also note that because the ID numbers of the citizens are reassigned between stages, there can be no equilibria in which the authority conditions her targeting strategy in Stage 2 on a specific citizen’s contribution choice in Stage 1.

The following Lemma describes the conditions under which a citizen will face the strongest incentives to contribute in the second stage.
Lemma 1 The incentives for citizen \(i\) to contribute in stage 2 are maximal when (1) \(a^2 = 1\); (2) \(t^2 = i\) if and only if \(\sigma_i^2 = 0\); and (3) \(c^2_j = 1\) \(\forall j \neq i\).

Informally, subjects face the strongest incentives to contribute when the authority invests in high accuracy information and targets a citizen for punishment if and only the signal associated with that citizen is incriminating; and when all other citizens contribute. If the authority does not invest or invests with some probability less than one, or targets a citizen for whom she received an exculpatory signal, or neglects to sanction in the presence of an incriminating signal, this reduces the likelihood of sanction for non-contributors and increases the probability of sanction for contributors, thus undermining deterrence. If other citizens do not contribute or contribute with some probability less than one, this reduces the likelihood of sanction for non-contributors, also undermining deterrence.

Lemma 1 permits the next result.

Lemma 2 Given the parameter values for both experiments, contributing to the public good in the either stage is a strictly dominated strategy.

Proof. Given Lemma 1, it is sufficient to demonstrate that not contributing is a best response when the incentives to contribute are maximal. The expected utility to the citizen of contributing given that all other citizens contribute and the authority invests in high quality information and targets optimally is given by

\[
E[u_i(c^2 = 1|c^2_j = 1\forall j \neq i, a = 1)] = Nry - (1 - p)f \sum_{k=0}^{N-1} \binom{N-1}{k} p^{N-1-k}(1-p)^k \frac{k+1}{k+1},
\]

while the expected utility to the citizen of not contributing is given by

\[
E[u_i(c^2 = 1|c^2_j = 1\forall j \neq i, a = 1)] = y + (N - 1)ry - pf \sum_{k=0}^{N-1} \binom{N-1}{k} p^{N-1-k}(1-p)^k \frac{k+1}{k+1}.
\]

Comparing these two values, citizen \(i\) will contribute if and only if

\[
(2p - 1)f \sum_{k=0}^{N-1} \binom{N-1}{k} p^{N-1-k}(1-p)^k > (1 - r)y.
\]

Substituting the experimental parameter values from above, the left side of the inequality simplifies to 8.856 and the right side to 12. Therefore, non-contribution is a best response when
incentives to contribute are maximal. If non-contribution is a best response when incentives to contribute are maximal, it must also be a best response when they are not. Therefore, contribution in the second stage is strictly dominated.

To demonstrate that not contributing is a strictly dominant strategy in the first stage, it is sufficient to demonstrate that a targeting strategy that employed the sanctioning authority in both stages to incentivize compliance in the first stage would not provide sufficiently strong incentives to make contribution a best response. Given that no citizen contributes in the second stage, and given random reassignment of ID numbers, the incentives for citizen $i$ to contribute in the first stage are maximized when $a^1 = 1; (2) t^1 = i$ if and only if $\sigma^1_i = 0; (3) c^1_j = 1 \forall j \neq i; and (4) t^2 = \emptyset$ if and only if $\sigma^1_i = 1 \forall i$. The probability that condition (4) is met given that conditions (1) through (3) are met is

$$\Pr(\sigma^1_i = 1 \forall i | \cdot) = p(1)^N,$$

if all citizens contribute, and $p(1)^{N-1}(1 - p(1))$ if all citizens but $i$ contribute. Because no citizen contributes in the second stage, each has an equal probability of being sanctioned if at least one citizen is observed not to have contributed in the first stage. The additional benefit of contributing in the first stage brought about by the prospect of group punishment in the second stage is therefore given by

$$\frac{(p(1)^N - p(1)^{N-1}(1 - p(1)))f}{N} = 1.536,$$

and from the calculations above, contribution in the first stage would be a best response if and only if $8.856 + 1.536 > 12$. The falsity of this inequality implies that non-contribution in the first stage is strictly dominant.

The next proposition gives a complete characterization of subgame perfect Nash equilibrium in the game.

**Proposition 1 (Equilibrium)** In a subgame perfect Nash equilibrium to the Study 1 game, (a) No citizen contributes in either stage; (b) the authority does not invest in high accuracy information in any period; and (c) the authority chooses any targeting strategy. In a subgame perfect Nash equilibrium to the Study 2 game, (a) No citizen contributes in either stage; and (b) the authority
chooses any targeting strategy.

**Proof.** The game may be solved by backward induction. After the second stage investment choice, the authority’s targeting choice is costless and offers no prospective benefit. Therefore, any targeting strategy is sequentially rational. Because the authority’s stage 2 investment choice is costly and offers no prospective benefit, the authority strictly prefers not to invest. By Lemma 2, citizens do not contribute. The same logic governs the first stage choices. ■

Because this is a complete information game, there is no private information for the authority to signal through his or her first stage investment choice. Suppose, however, that authorities possessed some extrinsic private information about, for example, their willingness to incur the cost of investment in the second stage, or their resoluteness. If this were the case, then the authority might use the first stage investment as a costly signal of his or her type. Then an observed relationship between first stage investment and second stage contributions could be an artifact of rational updating on the part of the citizens. As the following remark makes clear, however, signaling of this sort cannot be sustained in equilibrium.

**Remark 1 (No Separating Equilibrium)** Suppose the authority possesses private information extrinsic to the game about her type. There is no equilibrium in which this information is revealed, either partially or fully, by the authority’s investment choice in the first stage.

**Proof.** Suppose there is a separating equilibrium in which authorities for whom investment in high accuracy information and optimal targeting are best responses in equilibrium. By Lemma 2, even citizens fully informed that they were facing such an authority would strictly prefer not to contribute. But then such authorities would have no incentive to incur the cost of the first stage investment to reveal their types, a contradiction. ■