Identifying Legitimacy:

Experimental Evidence on Compliance with Authority

Eric S. Dickson, Sanford C. Gordon, and Gregory A. Huber[1]

[1]Dickson: Department of Politics, New York University, 19 W. 4th Street, New York, NY 10012, eric.dickson@nyu.edu. Gordon: Department of Politics, New York University, 19 W. 4th Street, New York, NY 10012, sanford.gordon@nyu.edu. Huber: Institution for Social and Policy Studies and Department of Political Science, Yale University, PO Box 208209, New Haven CT 06520, gregory.huber@yale.edu. We thanks seminar participants at Caltech, Northwestern, Princeton, Stanford, Stony Brook, UC Davis, UC Berkeley, the University of Nottingham, the University of Pennsylvania, the University of Virginia, Yale, the Sixth Biennial Meeting of the Social Dilemmas Working Group at Brown, the EGAP 15 conference at Rice, and the 2017 Washington Area Political Economy Conference at Georgetown; as well as John Bullock, Alessandra Casella, Pedro Dal Bó, Dimitri Landa, and Ken Shotts for comments and feedback. Vivek Ashok and Sebastian Thieme provided invaluable research assistance.

Abstract

Legitimacy is a foundational concern in political theory and across the social sciences. We consider the extent to which individuals' perceptions of an authority's legitimacy affect their intrinsic compliance motivations. Identifying this relationship empirically is challenging because actions and institutions that might affect legitimacy generally also affect citizens' material incentives. We formalize this identification problem and describe sufficient conditions under which a "personal" legitimacy effect can be separately identified. We then describe results of two public goods experiments that circumvent barriers to inference highlighted by our model. The first reveals a large, highly significant personal legitimacy effect. The second demonstrates that this effect is not reducible to subjects' motivation to "pay back" authorities. Our model and experimental results provide analytical coherence to the empirical study of legitimate authority while also providing valid causal evidence for its effects.

This paper considers how an individual's decision to comply with rules and behavioral norms are shaped by choices of authorities that enhance their legitimacy. The normative question of what constitutes legitimate authority has preoccupied political and legal philosophers at least since Plato, but concern regarding its empirical antecedents spans the social sciences. Because unlimited coercive capacity is prohibitively costly or otherwise infeasible, governing authorities who wish to facilitate social order must rely at least in part on a widespread consensus that complying with their edicts is "the right thing to do" (North, 1981; Levi, 1988).

Beyond that, however, social scientists, philosophers, and policymakers care about legitimate authority because they care about fairness, with the understanding that perceptions of unfairness can undermine the social compact. A series of recent incidents around the United States involving aggressive police tactics, particularly in communities of color, have undermined trust in the police (Department of Justice, 2015; Pew Research Center, 2016) and brought these issues to the forefront of national political consciousness. An improved understanding of how authorities can effectively enforce the law on the one hand, while simultaneously treating citizens fairly and maintaining legitimacy on the other hand, could yield substantial benefits for citizen well-being.

To construct analytically-grounded account of the antecedents of legitimate authority, we begin with a definition from social psychology: Tyler (1997) defines legitimacy as a "judgment by group members that they ought to voluntarily obey social rules and authorities *irrespective of the likelihood of reward or punishment*" (emphasis ours; see also French and Raven 1959). In other words, legitimacy obtains when the features of authoritative institutions, or the choices of individuals in positions of authority, enhance the *intrinsic* motivations of citizens to carry out certain social duties.

Empirically identifying legitimacy and its behavioral and procedural antecedents is challenging because of a fundamental problem of causal attribution. The problem arises because procedures or choices that enhance an authority's legitimacy may simultaneously alter cit-

izens' beliefs about the authority's capacity to bestow rewards and punishments. These extrinsic factors may, in turn, affect citizen behavior, quite apart from considerations of the authority's legitimacy.[1]

To sharpen the point somewhat, suppose a police officer behaves in a visibly "fair" manner, being scrupulous in her attention to detail and especially courteous to the suspects she interviews, and cultivating relationships with neighborhood citizens and business owners. If citizens on her beat obey the law more than they do a few blocks over, is it because they perceive her as legitimate, or because they perceive her as competent and informed and thus less error-prone? If her actions reduce the likelihood that she makes mistakes (arresting the innocent or failing to arrest the guilty), that will improve compliance through citizens' extrinsic motivations. But the relationship between her fair actions and citizens' compliance may also be driven by a positive affect toward the officer herself and the institutions she represents, which enhances citizens' intrinsic motivations to obey the law.

Or, consider the following example from the Department of Justice report on the Ferguson, Missouri police force (Department of Justice, 2015). According to the report, "City, police, and court officials for years have worked in concert to maximize revenue at every stage of the enforcement process, beginning with how fines and fine enforcement processes are established" (p. 10). This, when combined with extant racial disparities in policing, severely degraded the citizenry's trust in the police. Now suppose that citizens in Ferguson, as a consequence of mistreatment, are less likely to comply with the law's edicts than those in a similar town with a different law enforcement culture. A legitimacy-based account attributes this to the fact that citizens in Ferguson feel no psychological predisposition toward, or attachment to, the authority. But a deterrence-based account might attribute relative noncompliance to the fact that the police in Ferguson were less interested in appropriately punishing the guilty than they were in punishing as many people as possible, undermining material motives to comply.

[1]For a related critique of the analytical utility of the concept of legitimacy in the sociology of law, see Hyde (1983).

To develop a research design that can circumvent this problem, we describe a model in which agents are motivated to comply with an authority not only extrinsically by material rewards, but also intrinsically by their positive perceptions of the authority herself ("personal" legitimacy) and the underlying governance structure ("institutional" legitimacy). The authority can take a costly action that influences those perceptions, but may also affect the agents' extrinsic reward structure.

The model suggests that institutional legitimacy effects are generally not separately identifiable from the effects of extrinsic rewards and punishments, but also reveals sufficient conditions under which identification of a personal legitimacy effect is feasible. These conditions correspond to situations in which costlier actions on the part of the authority correspond to "better" probability distributions over governance structures. Under such conditions, an informative equilibrium exists in which authorities who care more about, e.g., the fairness of governance undertake costlier actions, and citizens observe and act based on variation in those actions while holding material incentives fixed.

We then describe the design, implementation, and analysis of two novel public goods experiments. The structure of these experiments relates them to experimental research on cooperation with third party punishment in different institutional contexts. Closest to our approach are papers by Dal Bó, Foster, and Putterman (2010), Baldassarri and Grossman (2011), and Grossman and Baldassarri (2012), who adopt an experimental approach to consider whether electing a centralized sanctioning authority can improve cooperation. The experiments permit us to isolate a personal legitimacy effect induced by a central authority's costly attempt to improve an enforcement mechanism's accuracy in assigning punishments to non-contributors (i.e., its procedural fairness).

We find strong evidence of a personal legitimacy effect: specifically, the authority's mere attempt to implement a fairer procedure increases the probability a citizen contributes by 10 to 12 percentage points. Thus, we provide direct behavioral evidence in support of the argument that legitimacy's normative core has positive behavioral implications. Furthermore,

our approach allows us to compare the magnitude of these personal legitimacy effects to the effects of institutional quality itself.

In the second experiment, we dig deeper into the mechanism underlying the personal legitimacy effect by introducing an additional random element: after the authority chooses an investment and an enforcement strategy, but before citizens choose whether or not to contribute, a coin flip that is observed by the citizens determines whether the authority will materially benefit from the public good. We can therefore assess whether personal legitimacy operates by activating a sense of reciprocity toward authorities among citizens, or through an intrinsic motivation that operates independently of reciprocity. We find that the personal legitimacy effect is no smaller when authorities do not benefit materially from the public good than when they do. Thus, the personal legitimacy effect must be understood as a motivation that exists over and above a feeling of debt to the authority.

The question of what makes authority legitimate has occupied empirical social scientists since Weber's (1947; 1978) seminal writings on the subject in the early 20th century. More recently, the prominent behavioral research on the subject has emerged at the intersection of social psychology (e.g, Tyler, 1990; Tyler and Ho, 2002) and criminal justice (e.g., Bottoms and Tankebe, 2012; Hough et al., 2010), with particular attention paid to the relationship between perceptions of procedural justice (Lind and Tyler, 1988) and legitimacy. Political science research on legitimate authority has paid closest attention to attitudes on legitimacy and the closely related notion of "diffuse support" (Easton, 1965), with a particular focus on support for courts (e.g., Caldeira and Gibson, 1992; Gibson, Caldeira, and Baird, 1998; Gibson, Caldeira, and Spence, 2003; Gibson, 2008). Others have examined the legitimacy of international organizations (Barnett, 1997; Hurd, 1999; Grant and Keohane, 2005). The common thread in these two political science research traditions is the question of how institutions with limited or no enforcement power achieve deference in spite of their politically tenuous position.[2]

[2]A related research agenda considers deference to autonomous bureaucrats; see, esp., Carpenter 2002. Note that the question of institutional legitimacy is related to, but analytically

One challenge that each of these literatures faces is the difficulty of isolating legitimacy-based from other, possibly extrinsic, motivations to support institutions. Thus, a central contribution of our research is to provide a novel analytical framework and empirical approach for disentangling extrinsic and intrinsic motivations to comply with authoritative institutions. In the conclusion, we discuss the broader implications of isolating these effects for institutional and policy design.

1 A Model of Legitimate Authority

We begin by formalizing the intuition that citizens' perceptions of the legitimacy of an authority or governing institutions can enhance intrinsic compliance motivations, and that actions by an authority can affect those perceptions. The model, and the experiments that follow, are in keeping with the longstanding practice in experimental social science of using contributions in a public goods game as a metaphor for social norm compliance (Ledyard, 1995). To keep the analysis as simple as possible, we abstract away from systematic differences among citizens that might induce distributive conflict among them. In such circumstances, citizens may differ among themselves in their evaluations of an institution's or authority's legitimacy.[3] Instead, we restrict our attention to *procedural fairness*, which social psychologists have identified as an important source of an authority's legitimacy (e.g., Tyler, 1990; Paternoster et al., 1997; Murphy, 2005). We operationalize procedural fairness as the accuracy with which a governance structure assesses compliance and noncompliance (and metes out penalties). Ex ante, all citizens will prefer that an authority exert greater effort to achieve higher accuracy. But because such effort is costly, authorities may not. More accurate institutions may be perceived as more legitimate, but in line with the causal attribution problem described in the introduction, will also, via a standard deterrence logic,

distinct from, institutional trust, as an individual might trust an institution owing to an assessment of its material performance.

[3]We leave the treatment of such situations as an avenue for future research.

affect the material incentives of citizens to comply.

To fix ideas, assume an authority and n citizens indexed by $i = 1, \ldots, n$, each with corresponding endowment normalized to one. Each citizen makes a binary choice $c_i \in \{0, 1\}$ to contribute the endowment to a public good or keep it for him or herself. Total contributions to the public good are given by $C = c_i + C_{-i}$, where $C_{-i} = \sum_{j \neq i} c_j$. We will refer to citizens who contribute as *compliant*, and those that withhold contribution as *noncompliant*. Citizens benefit from the return on the public good, $r(C)$, plus an intrinsic benefit $w(\cdot)$ from contributing, which we decompose in greater detail below.

The authority exerts costly effort $e \in \mathbb{R}^+$ that affects the *quality* of an enforcement mechanism, $q := \mathbb{R}^+ \to (\frac{1}{2}, 1)$. We assume that $q(e)$ is nondecreasing in e. In the current context, quality refers to accuracy: the probability that a sanction is correctly applied to compliant citizens and withheld from noncompliant citizens. The authority is characterized by a type, β, which is her own private information, and which scales the extent to which she primitively prefers a more procedurally fair institution, *independent of its effect on contributions to the public good*. For example, the authority may loathe the prospect of "punishing the innocent" even holding constant the effect punishing the innocent has on contributions.[4] The authority's utility is simply $U_A = r(C) + \beta q(e) - e$.

The sequence of the game is as follows:

1. The authority observes her own type, β, and then chooses effort e.

2. Citizens observe e and q, and then simultaneously choose whether to contribute c_i.

3. Enforcement mechanism implemented and payoffs realized.

Each citizen's intrinsic benefit from contributing, $w(\gamma_i, q(e), \hat{\beta}(e))$, is positively affected by three factors. The first is an idiosyncratic motivation to contribute, γ_i, which is drawn

[4]We focus on this concern for institutional quality specifically because social psychologists have hypothesized it to be a core source of legitimacy. That being said, similar results obtain if the authority's type refers to the extent to which she cares about the welfare of citizens, or to which she minds exerting costly effort. What is critical to the mechanism is simply that citizens' perceptions of the authority's type enhance their intrinsic motivation to contribute.

from a continuous distribution with cdf $G(\cdot)$. The second is *institutional legitimacy*: the extent to which a higher quality institution engenders greater affinity toward it and hence toward the act of contribution. The third term is the authority's *personal legitimacy* in the eyes of the citizen: the citizen's posterior belief about the extent to which the leader cares about quality/fairness.

We indicate the dependence of $\hat{\beta}$ on e, noting that at present, e fully determines q and that in any pooling equilibrium, $\frac{\partial \hat{\beta}}{\partial e} = 0$.

In what follows, we will assume that $r(C) = \rho C$, where $\rho \in (\frac{1}{n}, 1)$ is the marginal per capita rate of return on the public good. The expected utility to the citizen of contributing and not contributing are, respectively,

$$\begin{aligned} E[U_i(c_i = 1)] &= \rho(1 + E[C_{-i}]) + w(\gamma_i, q(e), \hat{\beta}(e)) - (1 - q(e))s \\ E[U_i(c_i = 0)] &= 1 + \rho E[C_{-i}] - q(e)s, \end{aligned}$$

where s is an exogenous sanction level. Comparing these quantities, citizen i will contribute if and only if

$$w(\gamma_i, q(e), \hat{\beta}(e)) \geq (1 - \rho) - (2q(e) - 1)s. \tag{1}$$

Given monotonicity of $w(\cdot)$ in γ, the citizen's decision problem will be characterized by a cutpoint, γ^*, with the citizen contributing if and only if $\gamma_i \geq \gamma^*$, where γ^* is defined implicitly by (1) at equality. The probability that any individual citizen contributes is then given by $1 - G(\gamma^*)$.

The foregoing serves to elucidate the following fundamental identification issues inherent to the empirical study of legitimacy.

A. Deterrent and institutional legitimacy effects of quality are not generally separately identified.

From above, the probability a citizen contributes is monotonically decreasing in γ^*. Implicitly differentiating equation (1) at equality and rearranging yields

$$\frac{\partial \gamma^*}{\partial q} = -\left(2s + \frac{\partial w}{\partial q}\right) \Big/ \left(\frac{\partial w}{\partial \gamma}\right).$$

As is evident from the expression, an increase in q simultaneously increases the potency of the material incentives to contribute (via the $2s$ term) while also increasing the positive affect toward contributing through the intrinsic incentive channel (via $\frac{\partial w}{\partial q}$).

This result has two immediate implications. First, it is a truism that large numbers of people comply with the law even if the chance of detected noncompliance is minuscule and the penalties slight. We would anticipate, even holding material incentives and idiosyncratic motivations constant, greater compliance under an institutional apparatus perceived by citizens as higher quality – via the institutional legitimacy channel.

The second implication concerns causal attribution: attempts to isolate a legitimacy effect by regressing compliance (or stated willingness to comply) on institutional quality (or perceptions of quality) will be confounded by the extrinsic compliance incentive. Note that *simple randomization of the treatment does not resolve this problem*: randomization, if feasible, would permit a valid causal estimate of the total effect of institutional quality, but not the effect attributable solely to institutional legitimacy.

In the social science literature, there are two standard empirical approaches to estimating the legitimating effect of procedurally fair institutions. The first is survey-based: query citizens who have had encounters with the law or another authority about whether they were treated fairly, whether they perceive the institution as legitimate, and whether they intend to comply with the law in the future. This is the approach taken by Tyler (1990), who 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), these studies are susceptible to potential problems of endogeneity and measurement error. The endogeneity problem arises because perceptions of fairness may be correlated with other unobserved individual- and institutional-level factors that also explain compliance. In the absence of a complete accounting of all differences between treatment and control groups (e.g., between those who experience more or less legitimate policing), the correlation between perceptions of legitimacy and compliance may be biased by factors associated both with policing and rates of compliance, including material incentives. Measurement error also poses a threat to inference because citizens who report perceptions of less legitimacy may misrepresent their behavior, and even random measurement error affecting a control variable will cause regression analysis to place weight on a "treatment" variable correlated with that imperfectly measured confound.

The second approach seeks to experimentally manipulate fairness: for example, Vermunt et al. (1996) and van den Bos (2001) assign subjects to treatments in which an authority evaluates an exam by grading one or more than one answer. This research, however, faces a different measurement issue: the measured outcome is not compliance with the authority per se, but affect toward an institution. Additionally, our analysis suggests that even in the absence of this concern, this design remains subject to the core causal attribution problem described here.

B. If institutional quality responds smoothly to the authority's effort, personal legitimacy effects induced by that effort are not separately identified from *either* deterrent or institutional legitimacy effects.

Implicitly differentiating (1) at equality and rearranging yields

$$\frac{\partial \gamma^*}{\partial e} = -\left[\left(2s + \frac{\partial w}{\partial q}\right)\frac{\partial q}{\partial e} + \left(\frac{\partial w}{\partial \hat{\beta}}\right)\left(\frac{\partial \hat{\beta}}{\partial e}\right)\right] \Big/ \left(\frac{\partial w}{\partial \gamma}\right)$$

The above expression shows three channels through which the authority's potentially legitimating effort e propagates through the citizen's contribution choice. The first is the deterrence channel: higher e corresponds to higher q, which in turn corresponds to a greater material motivation to contribute. The second is the institutional legitimacy channel: higher e corresponds to higher q, which corresponds to higher institutional legitimacy and a greater intrinsic motivation to contribute. The third channel is the personal legitimacy channel: in any informative equilibrium, $\hat{\beta}(e_2) > \hat{\beta}(e_1)$ for any $e_2 > e_1$. Thus, a costlier action on the part of the authority results in a higher posterior belief on the part of the citizen about the authority's type, which, again, increases the intrinsic motivation to contribute.

Consequently, regressing contributions on a measure of the authority's costly action will not separately identify a personal legitimacy effect unless the first two channels can be otherwise held constant. Note that for the same reasons as above, randomization does not fix this issue either. But the futility of random assignment is in some sense even starker here: the channel through which e operates on personal legitimacy is the citizens' posterior beliefs about the authority's type in equilibrium. If e were randomly assigned, it would be uninformative with respect to type, and thus could not affect citizens' beliefs about the authority.

A necessary condition to recover an effect of the costly action e through the personal

legitimacy channel (given the assumption of positive responsiveness of the intrinsic benefit to q) is $\frac{\partial q}{\partial e} = 0$ at some q. In other words, the empirical challenge is to hold q constant while inducing variation in e. But this presents a dilemma. Note that the authority seeks to maximize her expected utility,

$$N\rho(1 - G(\gamma^*)) + \beta q(e) - e.$$

A necessary and sufficient condition for a separating equilibrium (in which authorities with higher β choose higher values of e) is that β and e are complements. As is clear from this expression, this condition is violated if $\frac{\partial q}{\partial e} = 0$ for that level of q, in which case authorities of higher type lack the incentive to spend more to distinguish themselves from authorities of lower type. The implication is that a condition that must hold in order to incentivize the authority to take an action that induces personal legitimacy effects must be violated in order to empirically identify those effects.

C. Identifying personal legitimacy effects is feasible if e induces lotteries over q with some overlapping support.

In the foregoing, the relationship between q and e was deterministic. Suppose that we break this relationship as follows: Let $\theta(q|e)$ represent the conditional density of q given the authority's effort e, and suppose that for any e_1, e_2 with $e_2 > e_1$, (1) $\theta(q|e_2)$ is "better" than $\theta(q|e_1)$ in the sense of first order stochastic dominance; and (2) both $\theta(q|e_2)$ and $\theta(q|e_1)$ have common support in a nonempty interval (q_1, q_2). The authority seeks to maximize

$$E_{q|e}\left[N\rho(1 - G(\gamma^*)) + \beta q\right] - e.$$

Under these assumptions, β and e remain complements, permitting separation on type. But the overlapping support of the conditional distributions implies that we can simultaneously

observe different actions taken by the authority coincident with constant (or near-constant) levels of institutional quality. These conditions permit us to separately detect a personal legitimacy effect.[5]

The implication is that an experiment seeking to recover the personal legitimacy effect can do so by creating stochastic variation in the linkage between the authority's behavior and the institutional environment that the behavior affects. Note also that this approach preserves the belief-updating mechanism fundamental to the personal legitimacy channel, and that "treatment assignment" of different values of e to different citizens comes not through manipulation by the researcher, but via the natural variation in extent-of-care (parameterized as β in the model) among the population of authorities.

2 The Baseline Experiment

2.1 Design

We designed an experiment that incorporates some of the insights described above to recover an estimate of the effect of an authority's personal legitimacy on citizen behavior. In the experiment, an authority chooses whether to try and improve the accuracy of the information that will be available in the assignment of punishments. The chief empirical innovation of the design is to sever the deterministic relationship between the authority's decision and the actual institutional environment that eventually results, which allows us to isolate the personal legitimacy effect as described above.

The setting of the experiment is a linear public goods game with a centralized authority: subjects choose whether or not to contribute to a public good and are subject to enforcement actions by another subject designated as an authority. The authority and her "citizens" share

[5]Analogously, in an observational setting, we could identify a personal legitimacy effect if we had a valid instrumental variable that, along with the authority's effort investment, induced variation in institutional quality, but (a) did not affect the effort investment itself and (b) only affected citizen behavior through the institutional quality channel.

a common interest in maximizing contributions to the public good. (This is analogous to a situation in which a police officer lives in the community he or she polices.) Although individual citizens would prefer to withhold their contributions irrespective of the behavior of other citizens, citizens would nonetheless be better off if all players contributed than if none did so.

Subjects interacted anonymously via networked computers. The experiments were programmed and conducted using the software z-Tree (Fischbacher, 2007). After giving informed consent according to human subjects protocols, 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 (30 tokens = US$1) in amounts determined by the outcomes of play. Subjects' overall payoffs in a given session were equal to the sum of payoffs from each of the 20 periods (converted into dollars), plus a US$7 show-up fee.

At the beginning of each period, subjects were each given an endowment of 20 tokens and randomly assigned to a group of five people, of which four were randomly assigned as citizens (Role A, in the neutral parlance of the experiment), and one as an authority (Role B). Group and role assignments were randomly reassigned after each period. In each period, individual group members in Role A were labeled with an ID number between 1 and 4, commonly known to be randomly assigned in each period. Each period consisted of one play of the following extensive form game:

1. Authority chooses to make a "big" (4 token) or "small" (0 token) investment in accuracy.

2. Accuracy level given authority's investment determined.

3. Authority learns realized accuracy level and chooses enforcement rule.

4. Each citizen learns authority's investment, realized accuracy level, and enforcement rule, and chooses whether or not to contribute endowment to common pot.

5. Signals generated and enforcement rule implemented; payoffs realized.

We now turn to a fuller description of key features of the design.

Accuracy. Depending on its authority's choice, each group would be assigned to one of three different accuracy levels. (1) Under "Low Accuracy Information," the signals generated about each individual citizen's contribution decision has a 40% error rate. This means that if a specific citizen in fact kept (allocated) his tokens, the computer would generate a signal that the citizen kept (allocated) his tokens with 60% probability, but would generate an incorrect signal that the citizen allocated (kept) his tokens with 40% probability. (2) Under "Medium Accuracy Information," the error rate is 25%. (3) Under "High Accuracy Information," the error rate is 10%.

If the authority chose the small investment, the group's realized level of accuracy would be low with 50% probability and medium with 50% probability. If the authority chose the big investment, the realized level of accuracy in the group would be medium with 50% probability and high with 50% probability. This instantiates the "overlapping lotteries" condition described in the previous section. The realized level of accuracy is revealed to the authority before her choice of enforcement rule, and to the citizens (along with the authority's investment choice and the enforcement rule) prior to their contribution choice.

Enforcement Rule. The enforcement rule chosen by the authority in stage 3 is carried out automatically in stage 5, and thus represents a binding commitment revealed to the citizens prior to their contribution choice. Authorities could select one of four enforcement rules:

- Deduct 24 tokens from each citizen for whom a signal of "kept" was generated

- Deduct 24 tokens from each citizen for whom a signal of "allocated" was generated

- Never deduct tokens from any citizen, irrespective of signals generated
- Deduct 24 tokens from all citizens, irrespective of signals generated

As a shorthand, we will refer throughout to the first of these rules as "PATS" (Punish According To Signal); the second as "anti-PATS;" the third as "never punish;" and the fourth as "always punish."

We will focus much of our attention on the large majority of cases in which the authority chose the PATS enforcement rule, a decision we explain in greater detail below. The function of the enforcement rule choice is to create conditions under which the subjects assigned to the role of citizen mentally associate the authority with the dispensing of punishment. Precommitment greatly simplifies the strategic problem for the citizens, by creating common knowledge about the administration of penalties. In the absence of precommitment, citizens might condition their contribution choices on posterior beliefs about the resoluteness of authorities based on the accuracy investment, in addition to higher order beliefs about the beliefs of other citizens.

Extrinsic Incentives. The marginal per capita rate of return (MPCR) for citizens and authorities alike is 0.4, meaning that for every 20-token contribution to the common pot, the authority and each citizen receive 8 tokens. The 24-token deduction described above is calibrated to make a citizen motivated purely by the material payoffs of the experiment indifferent between contributing and not contributing under medium accuracy and PATS. To see this, note that a citizen who does not contribute under PATS/Medium Accuracy keeps his or her 20 token endowment but has 24 tokens deducted as a punishment with .75 probability, yielding an expected payoff of $20 - 0.75 \times 24 + \hat{C}_{-i} = 2 + \hat{C}_{-i}$ tokens, where $\hat{C}_{-i}$ represents citizen i's beliefs about others' contributions. At the same time, a citizen who does contribute receives a return of $0.4 \times 20 = 8$ tokens from his own contribution, but has 24 tokens deducted as a punishment with only 0.25 probability, yielding an identical expected payoff of $8 - 0.25 \times 24 + \hat{C}_{-i} = 2 + \hat{C}_{-i}$ tokens. As described below, our identification strategy focuses on comparisons within the Medium Accuracy institution given the PATS

enforcement rule. The material indifference these parameter values induce is therefore useful because it should, a priori, maximize variation in citizen contributions in that setting, which will be critical for identifying legitimacy effects. Very strong or very weak material incentives to contribute would, by contrast, obscure legitimacy effects.

Citizens' incentives to contribute to the public good are higher under PATS than under alternative enforcement strategies, even in the presence of imperfect signals about the contribution decision. Accordingly, an authority motivated by marginal deterrence alone is always weakly better off selecting this enforcement rule.[6] Additionally, PATS is the only enforcement rule for which more accurate information improves the material incentives to comply. For these reasons, we focus the bulk of our analysis below on cases in which the authority chooses PATS.

Finally, because players are randomly assigned to new groups and roles at the end of each period and interact anonymously, they have no reason to condition their choices on behavior in past rounds, or in expectation of future actions or repeated interactions (e.g., cultivating reciprocity norms).[7]

2.2 Identification

At the moment citizens are choosing whether or not to contribute to the public good, they are fully informed about all factors affecting their extrinsic motivations to comply: all parameter values, the realized level of accuracy, and the enforcement rule. (Because contributions by other citizens are additively separable, they are differenced out in a purely material calculation of whether to contribute.) Additionally, they have also observed whether or not the authority made the big investment in accuracy. Conditional on the realized accuracy level, this choice by the authority is materially irrelevant. However, it may not be irrelevant

[6] In the presence of idiosyncratic shocks to citizens' motivations to contribute, PATS is in fact the only enforcement rule consistent with equilibrium play.

[7] Below, we consider a psychological notion of reciprocity between authority and citizen unrelated to repeated interactions.

Table 1: Identification of Legitimacy Effect, Holding Punishment Strategy Constant

		Realized Accuracy Level		
		Low	Medium	High
Authority Investment	Small	$\overline{C}_{S,L}$	$\overline{C}_{S,M}$	–
	Big	–	$\overline{C}_{B,M}$	$\overline{C}_{B,H}$

Cell entries denote sample average group contribution rates

to the personal legitimacy of the authority.

Given these features of the design, we are now in a position to make the experiment's instantiation of the identification strategy described in Section 2 explicit. Assume the authority has chosen the PATS enforcement rule. Citizens will then make their decisions in one of four circumstances, summarized in Table 1. Rows denote the authority's investment (Small or Big), while columns denote the realized accuracy level. Conditional on the authority's choice of Small Investment, the citizen is randomly assigned to Low or Medium accuracy. Conditional on the authority's choice of big investment, the random assignment is to Medium or High investment. The notation $\overline{C}_{r,k}$ denotes sample average group-level contribution levels in row r and column k.

An approach analogous to one conducted in much of the prior literature would be to pool rows and compare average contribution rates given the big and small investments. It is immediate, however, that this confounds the personal legitimacy, institutional legitimacy, and deterrence-based material motivations to contribute described above. The more appropriate comparison of interest is

$$\overline{C}_{B,M} - \overline{C}_{S,M},$$

the differences in contributions given Medium Accuracy between an authority who makes the big and small investment. Given the PATS enforcement rule, this holds all material

motivations for citizens to contribute constant. If citizens are motivated by institutional legitimacy concerns, those are held constant as well. The only difference between the two circumstances is that in one, the authority undertook a costly action to improve the institutional environment, while in the other she did not. Thus, this comparison yields a valid causal estimate of the personal legitimacy effect.

We will also consider two additional comparisons of interest. The first is $\overline{C}_{S,M} - \overline{C}_{S,L}$. This identifies the total effect (deterrence plus institutional legitimacy) of a change from Low to Medium accuracy, holding the authority's investment choice constant at Small. The second is $\overline{C}_{B,H} - \overline{C}_{B,M}$, which identifies the total effect of a change from Medium to High accuracy, holding the authority's choice constant at Big. Note that because we do not know the shape of the relationship between accuracy and institutional legitimacy, differences-in-differences will not yield an estimate of marginal deterrence fully purged of institutional legitimacy effects absent strong but untestable assumptions.

Of course, there are other factors in the experiment that may also affect citizen behavior. Individuals' experiences during earlier rounds may affect their later play, and their behavior may generally evolve over the course of the game. While our design in which individuals are randomized into different groups and roles over time and then interact anonymously seeks to prevent most sources of repeat-play dynamics, we nonetheless undertake a variety of approaches to account for such dynamics in the analysis that follows.[8]

2.3 Experimental Results

We conducted four experimental sessions at [Redacted] and two sessions at [Redacted]. Each of the 90 subjects who participated took part in one session only. At both institutions, participants signed up via a web-based recruitment system that draws on a large, pre-existing pool of potential student subjects. (Subjects were not recruited from the authors' courses,

[8]Note also that even if players condition their behavior on their prior group-level experiences, groups are reshuffled across rounds so those expectations should be identical in expectation across groups in subsequent rounds.

Table 2: Authority Choices in the Baseline Experiment: Group-Level Data

	Enforcement Rule				
	PATS	Anti-PATS	Never Punish	Always Punish	Total
Small Investment	81	7	40	18	146
Big Investment	185	8	15	6	214
Total	266	15	55	24	360

and did not receive course credit for participating.) 46% of the subjects were female, and the median age was 20. 8% of the subjects were Economics majors, though 27% majored in a social science department. Subjects earned an average of $22.91 (s.d. of $2.5), with a maximum of $30.10 and a minimum of $17.70. The average score on the quiz administered between the reading of the instructions and play of the experiment was 6.4 out of 8 (s.d. 1.6) with 60% of subjects receiving a score of 7 or higher, and 24 receiving a perfect score.

Authority Behavior. The data consist of 360 group-period interactions. Table 2 summarizes the authorities' investment and enforcement rule choices in those interactions. Authorities selected the "big" investment 214 out of 360 times, or 60%. There was a modest increase in investment over time: in the first five periods the average rate was 50%, and in the last five it was 64%. Among players who were in the authority role more than once, 41% choose each investment level at least once. As a consequence of these choices, in 21% of all group-periods the authority received low accuracy information, in 53% medium accuracy information, and in 27% high accuracy information. 97% of players experienced all three accuracy levels while in the citizen role and the remaining 3% experienced two.

Those acting in the authority role overwhelmingly chose the PATS (punish according to signal) enforcement rule (74% of the time), although a substantial minority chose to never punish (15%), and smaller proportions chose either to always punish (7%) or to punish according to the anti-PATS rule (4%).[9] The PATS rule is slightly less common in the first 5 periods of play than afterwards: PATS is chosen by 67% of players in the first five rounds

[9]Two players accounted for 47% of the cases of anti-PATS, while eight players chose the anti-PATS rule only once.

compared to 76% of the time in the remaining periods. Those who made the big investment in accuracy are more likely to choose PATS than those who made the small investment (87 versus 55% of the time), with those who made the small investment more likely to choose either to never (27 versus 7%) or always (12 versus 3%) punish. Because the investment decision affects the accuracy of the signals received by the authority, there is a similar relationship between realized accuracy levels and the enforcement rule.

Aggregate Citizen Contribution Behavior. Overall, citizens contributed their tokens to the public good 65% of the time, for an average group contribution rate of 2.6 out of 4 (median 3). Unlike in standard public goods games with no enforcement (e.g., Fehr and Gächter, 2000), contributions do not diminish over time. Figure 1 displays data on group-level contributions by period (data points are jittered for clarity), along with a local polynomial smoother. The average contribution rate rises slightly over time, from around 2.5 in the first five periods to 2.8 in the final five. 94% of players varied their contribution decisions, while the remaining players nearly evenly split between never and always contributing.

The figure masks considerable heterogeneity in the data, which we explore systematically below. Unsurprisingly, contribution levels are highest when the authority adopts the PATS enforcement rule. Under PATS, the average group contribution rate is 3.1 out of 4; under Never Punish, 1.0; under Always Punish, 1.4; and under anti-PATS, 0.8. (The median rate was 3 out of 4 under PATS and 1 out of 4 under the remaining enforcement rules.) The stark difference between citizen behavior under PATS and under the other enforcement rules is likely due in large part to the much greater likelihood of enforcement errors: aggregating over different accuracy levels, under PATS, non-contributors escaped punishment 32% of the time, and contributors were punished 22% of the time. By contrast, under all other enforcement rules and accuracy levels, non-contributors escaped punishment 72% of the time, while the corresponding figure for punishment of contributors was 42%.

Institutional Effects. Before considering our results on personal legitimacy, recall that

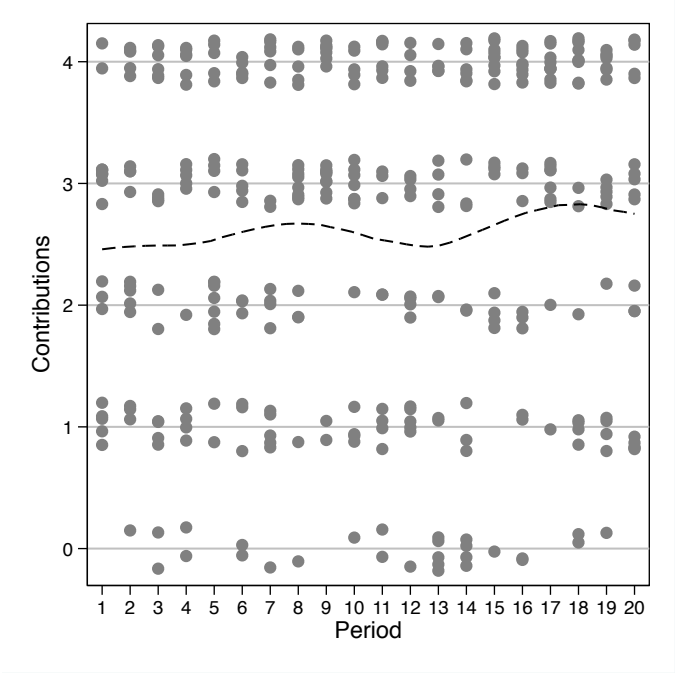

Figure 1: Group Contributions by Period
Note: Data jittered to enhance clarity of presentation. Dashed line is local polynomial smoother.

the experiment also permits us to identify two other separate quantities of interest: the total effect of a change from low to medium accuracy conditional on a small investment by the authority (and the PATS enforcement rule), and the total effect of a change from medium to high accuracy conditional on a big investment (and PATS). These effects correspond to the combined effect of the change in material incentives and institutional legitimacy associated with the changes in accuracy.

The relevant data are displayed graphically in Figure 2, which plots group average contributions by the authority's investment and realized accuracy level, conditional on the PATS enforcement rule. The first total institutional effect is obtained by comparing the pale gray column (bar #2, medium accuracy, small investment) with the white one (bar #1, low accuracy, small investment). The average group-level contribution rate in the former category was 2.93, as compared with 1.83 in the latter. The difference of 1.1 contributions is highly statistically significant and implies that the change in accuracy from low to medium induces

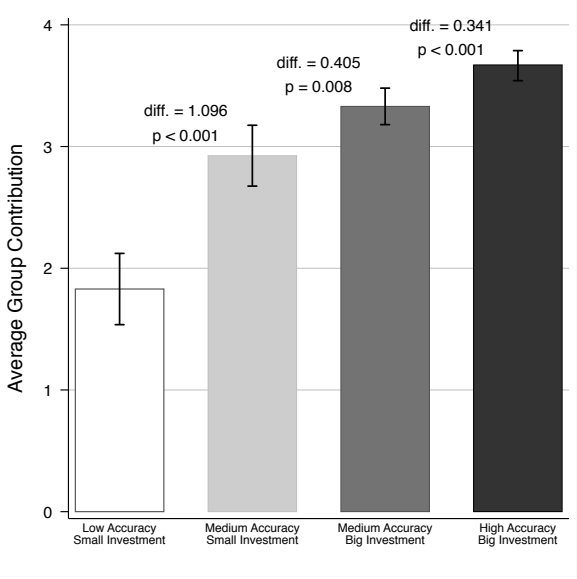

Figure 2: Group-level Contributions by Investment Decision and Accuracy Levels Given the Authority Choses Punish According to Signal

Note: Averages with bootstrapped 95% confidence intervals.

a 60% increase in contributions.

The second institutional effect is retrieved by comparing the black column (bar #4, high accuracy, big investment) with the dark gray one (bar #3, medium accuracy, big investment). Here, the effect is smaller, but still statistically significant: moving from medium to high accuracy conditional on the big investment is associated with an increase in group-level average contributions from 3.33 to 3.67, an increase of 10.2%.

The Personal Legitimacy Effect. We turn next to our analysis of the personal legitimacy effect, which is calculated as the different in contribution rates given the authority's big and small investment, holding constant accuracy (at medium) and enforcement rule (at PATS). Turning first to group-level average contributions, this calculation can be made comparing the light and dark gray bars (bar #3 and bar #4, respectively) in Figure 2. Under medium accuracy and a small investment, the average group-level contribution is 2.93 out of four. Under medium accuracy and a big investment, the group-level contribution is 3.33 on average. The difference in means, about 0.41, is significant at $p < 0.01$ (two-tailed), and

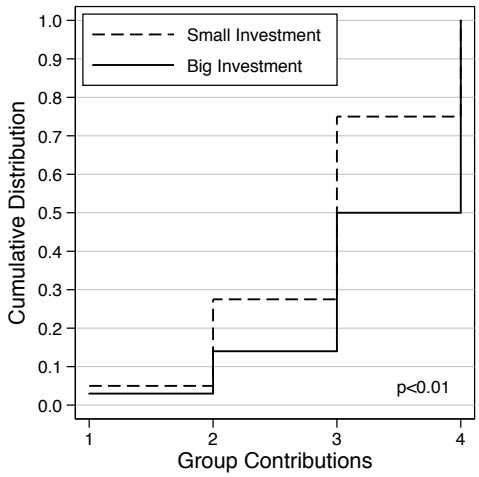

Figure 3: Empirical Cumulative Distribution Functions for Group Contribution Rates Given Medium Accuracy and PATS Enforcement Rule

Note: P-value is from Mann-Whitney tests.

corresponds to a 14% increase in the contribution rate.

We can go farther than a simple comparison of means by comparing the distributions of group-level contributions. Figure 3 displays empirical cumulative distribution functions for group-level contributions for the two comparison groups. The distributional plots confirm the inference from the sample averages: the distribution of outcomes given the big investment dominates that under the small investment (given medium accuracy and the PATS enforcement rule). For example, in half of the groups in which the authority made the big investment, all group members contributed; by contrast, only a quarter of groups in which the authority made the small investment enjoyed a 100% contribution rate. We also conducted a non-parametric Mann-Whitney test, with the null hypothesis that contribution rates under each condition were drawn from the same population. The test permits us to reject the null at $p < 0.01$.

Using simple raw descriptions of the data and leveraging the design features of our experiment, we have provided direct evidence of a personal legitimacy effect: authorities who

attempt to improve the quality of information on which they condition their punishments induce greater contribution levels by their group members than those who don't, holding the realized quality of information constant. Enhanced personal legitimacy therefore appears to be associated with real changes in citizen behavior.

Robustness. For the next part of the analysis, we allow for individual-level covariates and effects, shifting from a focus on group-level outcomes to the decisions of individual players in the citizen role. Specifically, we model individual i's decision whether to contribute in period t, $c_i \in \{0, 1\}$, as a function of the treatment variable (the authority's choice of the big investment in i's group g in time t, $a_{g,t} \in \{0, 1\}$), and additional covariates. Data are restricted to cases in which the authority chose the PATS enforcement rule and realized accuracy is medium. All specifications are OLS regressions with standard errors clustered at the group-period level (i.e., a group of four citizens in a single group in a single period), the level at which treatment is applied. Our baseline specification, then, is simply

$$c_{i,t} = \beta_0 + \beta_1 a_{g,t} + \gamma X_{i,t} + \varepsilon_{g,t}.$$

This framework allows us to understand whether our results are affected by accounting for players' experiences earlier in the game, the period of play, or other factors.

Table 3 shows these results. Note that because there are four citizens in a group, we should, in the absence of serious confounding, expect treatment effects estimates for the individual-level contribution analysis to be about 1/4 the 0.41 group-level difference in means described above. Column (1) is a simple OLS specification that mechanically demonstrates this: the effect of big investment in this specification is around 10 percentage points ($p < 0.01$, two-tailed).[10] In column (2), we include period indicators, which does not materially affect the estimated result. Column (3) adds each player's average experienced group contribution

[10]As noted in the text, this specification clusters at the period-group level. Both unclustered and robust standard errors are smaller. If we instead cluster at the session level, standard errors are slightly larger, with $p = .06$, two-tailed.

rate prior to the current period. Players who have experienced groups with more contributions in the past are more likely to contribute, but accounting for this effect modestly increases our estimate of the legitimating effect of the authority's investment in the current period. In the column (4) specification, we also account for each player's experience of how frequently the authority in prior periods made the big investment. The estimated legitimacy effect remains positive and statistically significant.

In the column (5) specification, we restrict our analysis to players who had already served as the authority in at least one prior period (recall that subjects are randomly reassigned to new roles in each period), in case the absence of such prior experience meaningfully affects the way in which subjects understand the implications of the investment choice. In this specification, we continue to find that individuals contribute more to the public good when the authority chooses a big investment. To reduce the possibility that our results are due to the behavior of subjects who did not understand the formal structure of the experiment, in column (6) we restrict our analysis to players who got at least 7 of the 8 quiz questions measuring subject comprehension correct prior to the beginning of the experiment.

Our most conservative analysis appears in columns (7) and (8). In these specifications, we include individual-level fixed effects. Estimating these models entails restricting the sample to players who experience medium accuracy and PATS in the citizen role following both a small and big investment by an authority. In the specification that does not include controls for period or past experience the estimate is 0.114 ($p < .01$). In the specification with these controls, it is 0.124 ($p < .01$). Thus, even after accounting for each player's individual propensity to contribute, we continue to find evidence that the authority's legitimating choice directly increased the likelihood a player contributed to the public good.

Although our design seeks to rule out the possibility that the authority's investment choice affects any material concerns once we account for the realized accuracy level and the chosen enforcement rule – that is, that the choice affects only the legitimacy channel we highlight – one possible violation of this assumption could take place if citizens form expec-

Table 3: Estimated Effect of Authority's Investment Choice, Given Medium Accuracy and PATS Enforcement Rule

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Authority Investment	0.101	0.109	0.115	0.125	0.118	0.111	0.104	0.124
(1=big, 0=small)	[0.038]	[0.037]	[0.037]	[0.038]	[0.048]	[0.052]	[0.039]	[0.046]
Average group contributions			0.072					0.078
prior to this period (0-4)			[0.030]					[0.064]
Average investment experience				-0.083				-0.047
prior to this period (0-1)				[0.081]				[0.144]
Constant	0.731	0.635	0.588	0.768	0.929	0.770	0.747	0.597
	[0.032]	[0.063]	[0.105]	[0.098]	[0.039]	[0.063]	[0.025]	[0.139]
Observations	560	560	540	540	404	338	447	428
R-squared	0.01	0.04	0.05	0.04	0.05	0.08	0.02	0.07
Number of unique subjects							65	65
Period fixed effects	N	Y	Y	Y	Y	Y	N	Y
Subject fixed effects	N	N	N	N	N	N	Y	Y

Dependent variable is player contribution decision (1=yes, 0=no).
OLS coefficients with group-period clustered standard errors in brackets.
Observations are individual contribution decisions given medium accuracy and PATS enforcement strategy.

tations about the behavior of other citizens' likely behavior on the basis of the authority's investment. For example, in period 1 those citizens who see an authority choose a big investment may infer that other members of their group are also more likely to be pro-social because the authority herself appears more pro-social. A related theoretical account argues that the leader's choice of investment is a coordinating device, signaling what all citizens should do.

We note that because the public good is linear in our design, there are no strategic complementarities that alter the incentives to contribute depending on other citizens' behaviors (as suggested in Levi 1988), so information about other players' types or their anticipated behaviors in a period should not increase contributions. Nonetheless, it is also reassuring that the specifications that account for average levels of previous group contributions and average prior authority behavior (i.e., columns (3) and (4)), which are direct measures of each citizen's prior experience and therefore proxies for each player's beliefs about the behavior of others, do not reduce the estimated legitimacy effect. Additionally, although the authority's choice is likely to provide novel information about other citizen's proclivities, we would expect the authority's choice to be more influential in earlier rounds, when players have less experience with other players' contribution behavior. Each period provides three pieces of evidence about those tendencies (the contribution choices of three fellow citizens in the group), compared to only a single authority investment choice. For this reason, it is notable that we find that the legitimacy effect is robust across subsets of the periods of the game. Figure 4 displays average contribution rates under medium accuracy and PATS, conditional on authority investment, over periods of play. Across all periods, average contributions are higher when the authority takes the legitimating action of the big investment.[11]

[11]The persistence of the legitimacy effect over time is also reassuring for another reason, in that it helps mitigate against explanations for our findings based on reciprocity across citizens. In standard public goods games without an enforcer, reciprocity has been shown to sustain cooperation when the shadow of the future looms large, but not in games (like ours) with anonymous interactions and random group reassignment. If reciprocity alone explained the pattern we observed, then we would expect it to be harder to sustain support for a big investment in later periods, when the game is about to end, than when there are

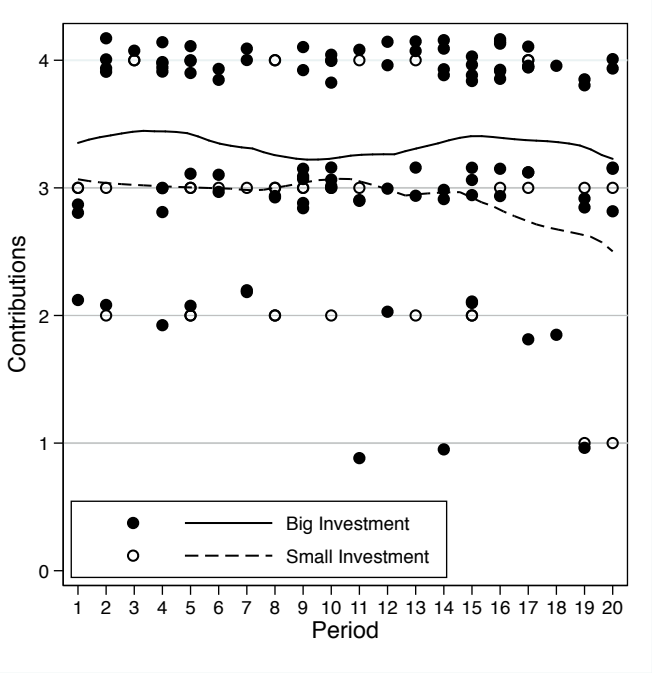

Figure 4: The Persistence of a Personal Legitimacy Effect Over Time
Note: Data jittered to enhance clarity of presentation. Observations are group contributions given medium accuracy and PATS enforcement strategy. Lines are local polynomial smoothers.

More formally, partitioning the data into the first ten or last ten periods and re-estimating the specification shown in column (2) yields an estimated legitimacy effect of .097 in periods 1-10 ($p = 0.01$, two-tailed) and .130 in periods 11-20 ($p = 0.06$, two-tailed).

3 Personal Legitimacy and Reciprocity Motivations

In the baseline experimental design, the authority benefits from citizen contributions to the public good. To the extent that higher accuracy induces greater rates of citizen contributions, this aspect of the design increases the authority's material incentive to undertake the big investment. However, this feature suggests that the personal legitimacy effect documented in the baseline experiment may be driven by a reciprocity motivation on the part of the subjects

many more rounds of play to follow. As the figure shows, however, we see no evidence that the legitimacy effect declines over time. Thus, it seems unlikely that this reciprocity account alone can explain the legitimacy result.

in their role as citizens: perhaps the increased rate of contribution reflects a desire by citizens to repay the authority for trying to do the right thing (cf., Akerlof, 1982). This reciprocity motive is analytically distinct from the mechanism modeled above, wherein beliefs about the authority's type affect the citizen's predisposition to contribute to the public good.[12]

3.1 Design and Identification

To assess whether this alternative psychological account explains our main result, we conducted a follow-on experiment that breaks the deterministic linkage between the citizen's contribution decision and the authority's material welfare. The design mimics the baseline experiment, with the key differences being that after the authority chooses an investment level and enforcement rule, an additional randomization takes place. With probability 0.5, the authority receives a flat fee of 20 tokens but does not benefit from contributions to the common pot; and with probability 0.5, she receives the 20 tokens plus 0.4 times the contributions to the common pot.

Importantly, this randomization takes place after the authority has made both of her decisions for the period but before the citizens have made their contribution decisions. Therefore, at the time the citizens are deciding whether to contribute to the public good, they both know all material factors that should affect their contribution decisions (as in the original experiment), as well as whether the authority's welfare will be affected by their contributions. When the authority directly benefits from contributions to the common pot, contributions are compatible with the reciprocity mechanism. By contrast, when the authority is paid only the flat fee, citizens know their contributions have no effect on the authority's welfare, and thus cannot be affected by these reciprocity concerns.

There is one other difference from the first experiment. For power reasons, there are two rather than three levels of accuracy: high (20% error rate) and low (40% error rate).

[12]Note that the psychological notion of reciprocity we consider here differs from the term's use in a repeated game framework.

Conditional on the small investment, accuracy was low with 75% probability and high with 25% probability; conditional on the big investment, those probabilities were reversed. Thus, we are able to estimate personal legitimacy for each compensation mechanisms under both low accuracy ($\overline{C}^m_{B,L} - \overline{C}^m_{S,L}$), modifying the notation from above with the superscript $m \in \{flat, benefit\}$) and high accuracy ($\overline{C}^m_{B,H} - \overline{C}^m_{S,H}$). Likewise, we can estimate the total effect of deterrence and institutional legitimacy under low and high investments, and under both levels of accuracy.

Note that in the current context, unlike in the baseline experiment, there is no realized level of accuracy that would make a subject motivated purely by extrinsic incentives indifferent with respect to contribution. Such a subject would strictly prefer to contribute given high accuracy, and strictly prefer not to given low accuracy. Given prior experimental results suggesting that subjects contribute more in public goods games than would be anticipated based on material incentives alone, ex ante we would expect greater variation in the contribution choice under low than high accuracy: under low accuracy, idiosyncratic motivations to contribute would push subjects closer to indifference, and under high accuracy, further from that indifference.

3.2 Experimental Results

We have data from 110 subjects gathered during 7 sessions conducted at [Redacted]. Subjects earned an average of \$19.27 (s.d. of \$2.5), with a maximum of \$25.67 and a minimum of \$12.47. The average score on the quiz administered between the instructions and the experiment was 6.15 out of 8 (s.d. 1.5), with 53% receiving 7 or higher and 17.3% receiving a perfect score.

Aggregate Authority and Citizen Behavior. Table 4 summarizes authority choices across all 440 group-period interactions. As in the first experiment, a majority of enforcers chose the PATS enforcement rule, although by a smaller margin (57% vs. 74%). This may reflect less concern for the welfare of the citizens in one's group given the lower expected

Table 4: Authority Choices in the Second Experiment: Group-Level Data

	Enforcement Rule				
	PATS	Anti-PATS	Never Punish	Always Punish	Total
Small Investment	142	35	82	35	294
Big Investment	108	5	13	20	146
Total	250	40	95	55	440

stakes of the choice in this setting. Additionally, fewer subjects in the role of the authority chose the big investment: 33% vs. 60%. Conditional on choosing the PATS enforcement rule, authorities chose the big investment 43% of the time (compared to 69% of the time in the baseline experiment). Overall, citizens contributed to the public good 48% of the time. This rate, while lower than in the first experiment, is similarly stable over time (See Figure 5).

Aggregated Institutional and Personal Legitimacy Effects. Before proceeding to our main analysis (disaggregating citizen behavior by the authority's compensation mechanism), we present summary data suggesting that the findings of the first experiment replicate in this alternative environment. Figure 6 displays average group contribution rates conditional on the PATS enforcement rule, for different authority investments and realized levels of accuracy.

The institutional effects are obtained by comparing, respectively, the dark gray (bar #3) and white (bar #1) bars (conditioning on small investment) and the black (bar #4) and pale gray (bar #2) bars (conditioning on big investment). They are unambiguous, and overwhelmingly statistically significant: conditional on the small investment, an increase from low to high accuracy nearly doubles the contribution rate, from 1.74 to 3.35. The institutional effect conditional on the big investment is smaller in magnitude, but still highly significant: an increase from 2.20 to 3.47, or 58%.

Next, we estimate the personal legitimacy effect, calculated at the group-period level and aggregating across authority compensation mechanisms. The effect conditional on high

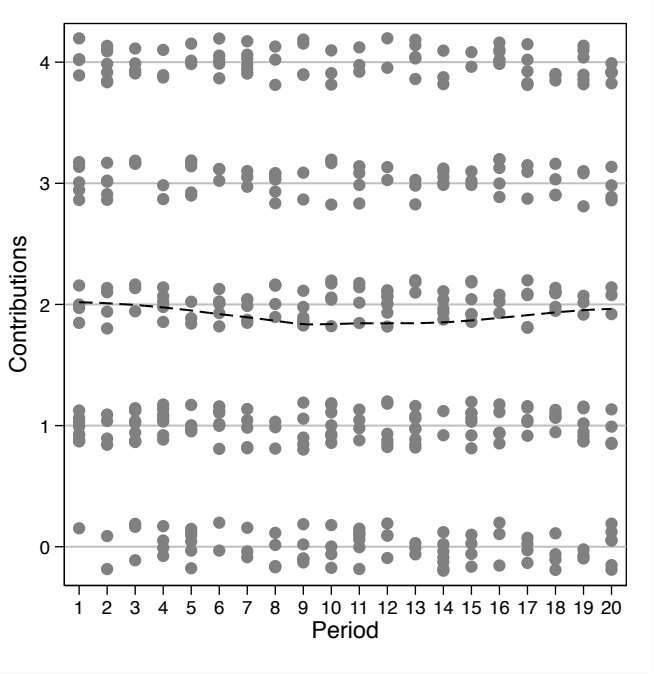

Figure 5: Group Contributions by Period, Second Experiment
Note: Data jittered to enhance clarity of presentation. Dashed line is local polynomial smoother.

accuracy is obtained by comparing the black and dark gray bars (bar #4 and bar #3, respectively) in Figure 6. We find the effect to be a statistically insignificant different of 0.12 contributions ($p = 0.39$, two-tailed). Recall that conditional on the high level of accuracy, subjects should strictly prefer to contribute to the public good based on extrinsic motivations alone, so the fact that average contribution rates are high under both levels of authority investment is unsurprising.

More relevant for the current discussion is the effect of authority investment conditional on the low accuracy level (comparing bar #2 and bar #1), recalling that extrinsic motives alone are insufficient to motivate contributions given this realization. Here, we observe a significant 27% increase in contributions, from an average of 1.74 to 2.2.

The Personal Legitimacy Effect and Authority Compensation. As above, our main analysis focuses on the effect on an individual citizen's contribution decision of the authority's investment choice in cases in which the authority chooses the PATS enforcement

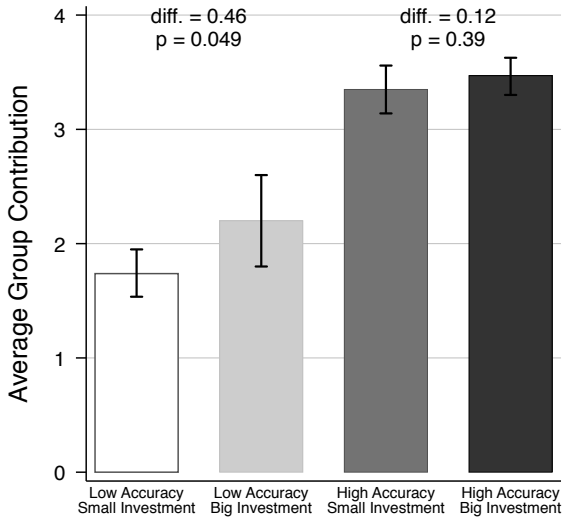

Figure 6: Group-level Contributions by Investment Decision and Accuracy Levels Given the Authority Choses Punish According to Signal, Second Experiment

Note: Averages with bootstrapped 95% confidence intervals.

rule, holding fixed the accuracy level. The added wrinkle is to condition this effect on the authority's compensation mechanism – which, recall, is realized after the authority's choices, but before the citizens'. Our analysis appears in Table 5.

Column (1) estimates the effect of the authority's investment decision by pooling across authority compensation mechanisms and accuracy levels. The specification suggests that the authority's choosing the big investment is associated with a statistically significant 6.6 percentage point increase in the likelihood of contribution. Confirming the results from the group-level analysis, the effect of accuracy is considerably larger, however, increasing contributions by 37 percentage points (relative to a baseline of 44%). Column (2) adds period-specific effects as well as two additional control variables capturing the subject's experiences in previous periods.

Column (3) disaggregates the personal legitimacy by the authority's compensation mechanism. If the reciprocity mechanism explains the earlier result, we would expect the coefficient on the interaction between the authority benefiting from the public good and the big investment to be positive while the coefficient on the authority's investment alone would diminish

Table 5: Conditional Effect of Authority's Investment Choice under Different Authority Compensation Mechanisms, Given PATS Enforcement Rule

	(1)	(2)	(3)	(4)	(5)	(6)
Accuracy	0.370	0.403	0.369	0.403	0.392	0.446
(1=high, 0=low)	[0.032]	[0.027]	[0.032]	[0.027]	[0.054]	[0.044]
Authority investment	0.066	0.063	0.076	0.058	0.123	0.116
(1=big, 0=small)	[0.031]	[0.028]	[0.041]	[0.035]	[0.075]	[0.059]
Average group contributions		0.154		0.154		0.154
in prior periods (0-4)		[0.023]		[0.023]		[0.023]
Average investments		0.001		0.001		-0.005
in prior periods (0-1)		[0.078]		[0.078]		[0.078]
Authority benefits			-0.017	-0.026	-0.025	-0.017
			[0.042]	[0.036]	[0.055]	[0.046]
Authority benefits × investment			-0.019	0.013	-0.007	-0.008
			[0.058]	[0.050]	[0.111]	[0.091]
Authority investment × accuracy					-0.073	-0.097
					[0.090]	[0.073]
Authority benefits × accuracy					0.026	-0.027
					[0.079]	[0.072]
Authority benefits × investment × accuracy					-0.032	0.038
					[0.132]	[0.113]
Constant	0.444	0.102	0.452	0.113	0.445	0.096
	[0.025]	[0.072]	[0.031]	[0.074]	[0.036]	[0.074]
Observations	1,000	960	1,000	960	1,000	960
R-squared	0.181	0.245	0.182	0.246	0.184	0.247
Period fixed effects	N	Y	N	Y	N	Y

Dependent variable is player contribution decision (1=yes, 0=no).
OLS coefficients with group-period clustered standard errors in brackets.
Observations are individual contribution decisions given PATS enforcement strategy.

in size. But this is not what we find. The estimates suggest that when the authority does *not* benefit from the public good, her choice of the big investment leads to a 7.6 percentage point increase in contributions ($p = 0.063$, two-tailed). When the authority does benefit, that rate decreases to 5.8 percentage points and is no longer statistically significant at conventional levels ($p = 0.19$). Column (4) adds the same vector of controls as in the column (2) specification, yielding an estimated effect of the authority's investment of 5.8 percentage points ($p = 0.10$) when the authority does not enjoy the public good, and a slightly larger effect of 7.1 percentage points when she does benefit ($p = 0.073$). In neither specification (3) nor (4) are the estimated effects of the authority's investment choice statistically distinguishable from one another.

Columns (5) and (6) disaggregate the effect of the authority's investment choice both by level of accuracy and authority compensation mechanism. The analysis is done via the inclusion of the triple interaction of investment choice, accuracy, and compensation mechanism as well as all first and second order terms. Per the column (5) specification, the authority's investment increases contributions by 12.3 percentage points ($p = 0.10$, two-tailed) given low accuracy and given the flat fee compensation mechanism for the authority. When the authority does benefit, and given low accuracy, this figure decreases in magnitude and statistical significance, to 11.5 percentage points ($p = 0.16$). As indicated in the group-level analysis, the effect of the authority's investment is quite small and indistinguishable from zero conditional on high accuracy. Column (6) adds the period dummies and past experience covariates; the estimates imply a comparable, though more precisely estimated ($p = 0.05$), effect of the authority's investment conditional on low accuracy and flat compensation, and comparable magnitudes and statistical significance under the other compensation mechanism and realized accuracy levels. In neither specification (5) nor (6) can we reject the null hypothesis that the effect of the authority's investment choice is the same regardless of the authority's compensation mechanism. This implies that even when reciprocity toward the authority is ruled out as an explanation, we continue to find evidence of a personal legitimacy

effect on subject contribution rates.

4 Discussion

This paper makes several contributions to the study of legitimate authority and its relationship with subordinate behavior. Theoretically, we distinguish between institutional and personal legitimacy and demonstrate the challenge of separately identifying these effects from one another as well as from material motivations for compliance, which persists even when institutions are randomly assigned. Empirically, the innovation of our experimental designs is to isolate the effect of an authority's enhanced personal legitimacy associated with *an effort* to secure a more procedurally fair institution (operationalized as a more accurate mapping of punishment to citizen culpability) from the legitimating and deterrence effects associated with the institutional itself. We do so by making the legitimating *action* of the authority probabilistically, rather than deterministically, related to the materially- and institutionally-relevant *consequences* of that action.

In our experiment, we find evidence consistent with the notion that an authority's personal legitimacy substantially increases the willingness of citizens to contribute to a collective good. This effect is robust to different sample restrictions and statistical modeling approaches. Additionally, in a follow-on experiment we show that this effect cannot be explained by an analytically distinct mechanism in which citizens "repay" an authority for her efforts to improve citizen welfare by isolating the authority's payoffs from those of the citizen. In this setting, in which the citizens' behavior has no material effect on the authority's well-being, we continue to find evidence supporting the personal legitimacy effect described in our model.

At the same time, we are cognizant that the stylized environment in which we are able to implement this design may depart from non-experimental settings in a number of respects. To underscore this point, consider two of the potentially artificial features of the experiments described above: that the authority can pre-commit to an enforcement rule, and that the

authority has capacity to punish all of the citizens in her group. The first of these features is particularly important because it means that each citizen in a group is operating with a common understanding of the environment of material incentives: they know the realized level of accuracy and the mapping between their behavior and the likelihood of punishment. In reality, of course, authorities tend not to pre-commit to enforcement strategies. In the absence of pre-commitment, an entirely different signaling account may operate, in which the citizens make inferences about likely enforcement strategies from the (ex ante) procedural investment. In such an environment, it might be quite challenging to disentangle the non-instrumental effect of the legitimating action from the instrumental consequences of the information the action conveys to the citizens.

Likewise, the assumption that the authority possesses the capacity to target all citizens is clearly unrealistic given that authorities typically have limited resources. In the context of the experiment, this is a valuable feature because it allows us to eliminate the strategic complementarities that arise in most enforcement environments. If an authority's capacity is limited, in many circumstances there will be multiple equilibria in which expectations about others' behavior come into play. If a citizen expects all other citizens to comply, non-compliance carries greater risks, because the probability of being targeted by the authority is higher; as such, the citizen's incentives to comply are heightened. If instead a citizen expects no or few other citizens to comply, non-compliance is less risky, because the probability of being targeted by the authority is reduced, weakening incentives for compliance. In an environment with strategic complementarities such as these, authorities' legitimating actions could affect citizen behavior by coordinating expectations about others' behaviors, thereby affecting equilibrium selection. Our design thus permits us to rule out explanations arising from these coordination issues.

This study departs from the corpus of prior research on procedural fairness and other sources of legitimacy in its use of an incentivized laboratory environment in which subjects receive financial compensation for their performance in a game. This approach offers several

important advantages. First, it permits us to establish a benchmark of rational behavior based on purely material motivations against which to compare actual behavior. With this benchmark in hand, 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.

Third, our specific experimental design also allows us to make progress in understanding the psychological origins of legitimacy. We show that leaders who attempt to obtain more procedurally fair institutions, operationalized here as those that are more accurate, have more personal legitimacy. Enhanced personal legitimacy, in turn, increases citizen compliance, and those effects are beyond those that arise due to changes in instrumental motivations or institutional legitimacy. Note that it is not realized procedures that we use to generate variation in legitimacy. Rather, we show that an authority's willingness to seek a better institution, independent of whether that procedure comes into being, enhances her personal legitimacy and alters citizen behavior. A necessary step for future work is to undertake additional research using different designs to understand whether the realization of fairer institutions also generates concomitant improvements in compliance via a legitimacy mechanism, although the methodological hurdles we identify here in disentangling material and non-material effects of institutions imply these efforts may be difficult.

To be sure, there are numerous potential sources of improved legitimacy aside from accuracy. For example, the authority may or may not take actions that appear biased against a member of a specific group. While the experiment described here focuses on accuracy, this design is sufficiently flexible to accommodate any number of other conceptions of sources of legitimacy. A task for future research is to estimate 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.

We conclude by noting the broader value of isolating different causal mechanisms associated with policies aimed at fostering cooperation and compliance. One might be tempted to argue that if some institution "works," it is irrelevant whether it does so owing to its intrinsic or extrinsic effects. But particularly given the frequency with which legitimacy is invoked causally, this risks serious misunderstanding. This misunderstanding may itself be undesirable, but if it in turn leads to bad policy advice, the consequences may also be deleterious from a public welfare perspective. In particular, there are many situations in which policymakers confront choices about which (costly) reform to implement. While it is often the case that enhanced legitimacy and improved material motivations for compliance move hand in hand, they do not always do so (e.g., in the case of policies surrounding racial profiling or other forms of group targeting) and sometimes interventions that affect one causal pathway more than another are differentially costly. Fully isolating and understanding the empirical consequences of legitimacy for compliance allows us to better make predictions and recommendations in situations in which the contours of policy involve choices along these lines.

References

Akerlof, George A. 1982. "Labor Contracts as Partial Gift Exchange." *The Quarterly Journal of Economics* 97 (4): 543–569.
URL: *http://www.jstor.org/stable/1885099*

Baldassarri, Delia, and Guy Grossman. 2011. "Centralized sanctioning and legitimate authority promote cooperation in humans." *Proceedings of the National Academy of Sciences* 108 (27): 11023–11027.
URL: *http://www.pnas.org/content/108/27/11023.abstract*

Barnett, Michael N. 1997. "Bringing in the New World Order: Liberalism, Legitimacy, and

the United Nations." *World Politics* 49 (4): 526–551.
URL: *http://www.jstor.org/stable/25054018*

Bottoms, Anthony, and Justice Tankebe. 2012. "Beyond Procedural Justice: A Dialogic Approach to Legitimacy in Criminal Justice." *Journal of Criminal Law and Criminology* 102 (1): 119–170.

Caldeira, Gregory A., and James L. Gibson. 1992. "The Etiology of Public Support for the Supreme Court." *American Journal of Political Science* 36 (3): 635–664.
URL: *http://www.jstor.org/stable/2111585*

Carpenter, Daniel P. 2002. *The Forging of Bureaucratic Autonomy*. Princeton, NJ: Princeton University Press.

Dal Bó, Pedro, Andrew Foster, and Louis Putterman. 2010. "Institutions and Behavior: Experimental Evidence on the Effects of Democracy." *American Economic Review* 100 (5): 2205–29.
URL: *http://www.aeaweb.org/articles?id=10.1257/aer.100.5.2205*

Department of Justice. 2015. "Investigation of the Ferguson Police Department.".

Easton, David. 1965. *A Systems Analysis of Political Life*. New York: Wiley.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *The American Economic Review* 90 (4): 980–994.
URL: *http://www.jstor.org/stable/117319*

Fischbacher, Urs. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics* 10 (2): 171–178.
URL: *https://doi.org/10.1007/s10683-006-9159-4*

French, John R., and Bertram Raven. 1959. "The Bases of Social Power." In *Studies in*

Social Power, ed. D. Cartright, and A. Zander. Ann Arbor, MI: University of Michigan Press pp. 150–167.

Gibson, James L. 2008. "Challenges to the Impartiality of State Supreme Courts: Legitimacy Theory and "New-Style" Judicial Campaigns." *The American Political Science Review* 102 (1): 59–75.
URL: *http://www.jstor.org/stable/27644498*

Gibson, James L., Gregory A. Caldeira, and Lester Kenyatta Spence. 2003. "Measuring Attitudes toward the United States Supreme Court." *American Journal of Political Science* 47 (2): 354–367.
URL: *http://www.jstor.org/stable/3186144*

Gibson, James L., Gregory A. Caldeira, and Vanessa A. Baird. 1998. "On the Legitimacy of National High Courts." *The American Political Science Review* 92 (2): 343–358.
URL: *http://www.jstor.org/stable/2585668*

Grant, Ruth W., and Robert O. Keohane. 2005. "Accountability and Abuses of Power in World Politics." *The American Political Science Review* 99 (1): 29–43.
URL: *http://www.jstor.org/stable/30038917*

Grossman, Guy, and Delia Baldassarri. 2012. "The Impact of Elections on Cooperation: Evidence from a Lab-in-the-Field Experiment in Uganda." *American Journal of Political Science* 56 (4): 964–985.
URL: *http://www.jstor.org/stable/23317168*

Hough, Mike, Jonathan Jackson, Ben Bradford, Andy Myhill, and Paul Quinton. 2010. "Procedural Justice, Trust, and Institutional Legitimacy." *Policing: A Journal of Policy and Practice* 4 (3): 203–210.
URL: *+ http://dx.doi.org/10.1093/police/paq027*

Hurd, Ian. 1999. "Legitimacy and Authority in International Politics." *International Organization* 53 (2): 379–408.

Hyde, Alan. 1983. "The Concept of Legitimation in the Sociology of Law." *Wisconsin Law Review* 1983: 379–426.

Ledyard, John. 1995. "Public Goods." In *Handbook of Experimental Economics*, ed. John H. Kagel, and Alvin E. Roth. Princeton, NJ: Princeton University Press pp. 111–194.

Levi, Margaret. 1988. *Of Rule and Revenue*. Berkeley, CA: University of California Press.

Lind, E. Allan, and Tom R. Tyler. 1988. *The Social Psychology of Procedural Justice*. New York: Plenum Press.

Murphy, Kristina. 2005. "Regulating More Effectively: The Relationship between Procedural Justice, Legitimacy, and Tax Non-compliance." *Journal of Law and Society* 32 (4): 562–589.
URL: *http://dx.doi.org/10.1111/j.1467-6478.2005.00338.x*

North, Douglas C. 1981. *Structure and Change in Economic History*. New York: W. W. Norton & Company.

Paternoster, Raymond, Robert Brame, Ronet Bachman, and Lawrence W. Sherman. 1997. "Do Fair Procedures Matter? The Effect of Procedural Justice on Spouse Assault." *Law and Society Review* 31 (1): 163–204.
URL: *http://www.jstor.org/stable/3054098*

Pew Research Center. 2016. "On Views of Race and Inequality, Blacks and Whites Are Worlds Apart.".
URL: *http://www.pewsocialtrends.org/2016/06/27/on-views-of-race-and-inequality-blacks-and-whites-are-worlds-apart/*

Tyler, Tom R. 1990. *Why People Obey the Law*. New Haven, CT: Yale University Press.

Tyler, Tom R. 1997. "The Psychology of Legitimacy: A Relational Perspective on Voluntary Deference to Authorities." *Personality and Social Psychology Review* 1 (4): 323–345.

Tyler, Tom R., and Yuen J. Ho. 2002. *Trust in the Law: Encouraging Public Cooperation with the Police and Courts*. New York: Russell Sage Foundation.

van den Bos, Kees. 2001. "Uncertainty Management: The Influence of Uncertainty Salience on Reactions to Perceived Procedural Fairness." *Journal of personality and social psychology* 80 (6): 931–941.

Vermunt, Riël, Arjaan Wit, Kees van den Bos, and E. Allan Lind. 1996. "The effects of unfair procedure on negative affect and protest." *Social Justice Research* 9 (2): 109–119. **URL:** *https://doi.org/10.1007/BF02198075*

Weber, Max. 1947. *The Theory of Social and Economic Organization*. New York: Simon and Schuster.

Weber, Max. 1978. *Economy and Society*. Berkeley, CA: University of California Press.