To what extent do individuals’ perceptions of legitimacy affect their intrinsic motivations to comply with an authority? Answering this question has critical implications for law enforcement, but is challenging because actions or institutions that affect intrinsic motivations typically also affect extrinsic, material ones. To disentangle these, we propose a novel experimental approach that separately identifies the effect of an authority’s costly action to improve enforcement fairness on citizen behavior through both intrinsic and extrinsic channels. In Experiment 1, the authority’s simple attempt to institute fairer enforcement increases pro-social behavior by 10-12 percentage points via the intrinsic channel. A follow-up experiment demonstrates that this is not motivated by citizen attempts to “pay back” authorities. Our findings provide causally credible evidence that an authority’s actions can directly shape citizens’ behavior by enhancing her legitimacy, and have important implications.
in policy domains where this conflicts with other incentives.

Teaser

Behavioral experiments disentangling material and psychological incentives show subjects comply more with legitimate authorities.

Introduction

This paper tests whether an individual’s decision to comply with rules and behavioral norms is shaped by perceptions of an authority’s 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, infeasible, or otherwise undesirable, 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” (1, 2).

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 (3) and brought these issues to the forefront of national political consciousness. An improved understanding of how authorities can effectively enforce the law while simultaneously treating citizens fairly and maintaining legitimacy could yield substantial benefits for citizen well-being.

To construct an analytically-grounded account of the antecedents of legitimate authority, we begin with a definition from social psychology: legitimacy is a “judgment by group members that they ought to voluntarily obey social rules and authorities irrespective of the likelihood
of reward or punishment’ (4) (emphasis added). In other words, legitimacy obtains when the features of authoritative institutions, or the choices of individuals in authority, enhance the intrinsic motivations of citizens to carry out certain social duties.

Empirically identifying how an authority’s actions can affect these intrinsic motivations is challenging because of a fundamental causal attribution problem: procedures or choices that enhance these motivations may simultaneously alter citizens’ 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.

This attribution problem is particularly pronounced in the realm of policing, where, for example, allegations of abuse invariably call into question not only the efficacy, but also the fairness – and thus legitimacy – of law enforcement. Suppose a police officer behaves in a visibly “fair” manner, being scrupulous in her attention to detail and courteous to the citizens with whom she interacts. 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 thus less error-prone? Both intrinsic and extrinsic motivations are in play.

Or, consider the following example from the Department of Justice report on the Ferguson, Missouri police force (5). 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...” (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 (to maximize revenue), thereby undermining material motives to comply.

We present an experimental paradigm in which the effects of an authority’s behavior on the material incentives of citizens are carefully delineated and decoupled from its effect on intrinsic motivations through a novel randomization. In our framework, an authority can take a costly action meant to improve the fairness of punishment for failure to contribute to a public good. However, the action is only effective with a known and predetermined probability. Citizens are made aware both of the authority’s choices as well as the actual material incentives they face, based on the randomly realized degree of fairness of the institution. By randomizing the citizen’s extrinsic incentives in this way, we are thus able to causally identify the effect of the authority’s action on citizen behavior via the intrinsic channel, holding constant the citizens’ material incentives.

Our paradigm, which employs a public goods contributions framework to proxy social cooperation as overseen by an authority, relates to experimental research on cooperation with third party punishment in different institutional contexts. Closest to our approach are (6–8), which adopt an experimental approach to consider whether electing a centralized sanctioning authority can improve cooperation. (Relatedly, in a recent field experiment in Vietnam, (9) found that the opportunity to participate in rulemaking improved corporate compliance and increased the esteem with which corporations view regulators.) Rather than focusing on the process of decision-making or leader selection, we instead examine the policy choice of an authority, restricting our attention to a situation in which an authority is not selected by the citizenry. This is analogous to many citizen-authority interactions such as between citizens and police officers, students and teachers, and prisoners and prison guards. Our experiments permit us to isolate a legitimating 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 fair-
ness). Procedural fairness is one of the key sources of legitimacy identified in prior research on the psychology of legitimacy (10).

Our work also contributes to a vibrant literature on centralized punishment in experimental economics. (11) and (12) study potential benefits of centralized punishment, but focus on the potential increases in efficiency that this institution can offer, especially relative to peer-to-peer punishment; this work, however, does not directly address the question of intrinsic motivations to comply with authority. As in this study, (13) explores a centralized punishment context in which the authority can engage in more or less meritorious behavior; however, they focus on very different measures (e.g., discrepancies between the authority’s punishment choices and his own contribution decisions) and do not distinguish between intrinsic and extrinsic motivations in shaping citizens’ future contributions. (14) and (15) explore the effects of democratically electing a punisher – rather than having one be exogenously appointed – on citizen contributions. This work is, like ours, framed in terms of studying intrinsic incentives, but the empirical results flowing from their paradigms are for the most part statistically insignificant. Finally, (16) demonstrates that leaders can improve group outcomes in a coordination game either through communication or through affecting material incentives; their framework is game-theoretically very different from ours, and they also do not attempt to measure intrinsic motivations directly.

In our main experiment, we find strong evidence that the authority’s fairness-seeking action markedly increases citizens’ willingness to engage in pro-social behavior, independent of the material consequences of that action. Specifically, in a sample of University students, the authority’s mere attempt to implement a fairer procedure increases the probability that a citizen contributes to the public good by 10 to 12 percentage points.

The effect we observe in the main experiment is consistent with several potential mechanisms: (a) a “normative legitimation” account in which the authority’s action primes a citizens’ sense of obligation to behave pro-socially; (b) a “conditional cooperation” account in which the
authority’s action conveys information about other citizens’ willingness to contribute; and (c) a “direct reciprocity” account in which the citizen seeks to “pay back” the authority for the costly investment.

Analysis of ancillary results from our main experiment cast doubt on the second mechanism. In a follow-up experiment, we consider whether the results from the main experiment are driven by reciprocity. The experiment accomplishes this by randomizing whether or not the authority materially benefits from the public good. We find that the nature of the authority’s compensation does not moderate the effect of the authority’s action. While we cannot eliminate all conceivable alternative mechanisms, this conjunction of findings establishes the preeminence of the normative legitimation account.

Conceptual Issues

Legitimate Authority: The Identification Problem In the social science literature, a standard approach to estimating the legitimating effect of procedurally fair institution or authority action is survey-based (17–19): query citizen \( i \) who has had an encounter with the law or another authority about whether or not they were treated fairly \( (A_i \in \{\text{unfair, fair}\}) \) and whether or not they intend to comply with the law in the future \( (C_i \in \{\text{don’t comply, comply}\}) \). Such studies interpret an association between variables \( A \) and \( C \) as evidence that it is subjects’ perceptions of the legitimacy of the legal process that affect their compliance intentions.

Fig. 1A depicts a directed acyclic graph (DAG) intended to illustrate two inferential issues with this approach. (Measurement error also poses a threat to inference, as citizens may misreport their perceptions or behavior.) The first is omitted variables bias: as noted in (20), perceptions of fairness may be correlated with numerous unobserved individual- and institutional-level factors that also explain compliance (denoted \( U \) in the figure). The other inferential challenge is that perceptions of fairness may affect compliance via the two channels described in the intro-
duction: an extrinsic channel in which $A$ affects material motivations $M$, which in turn affect $C$; and an intrinsic channel in which $A$’s actions directly affect $C$. Even absent unobservable confounds, an observed association between $A$ and $C$ cannot distinguish between these channels. Studies that randomize the fairness of an experimental treatment (20, 21) may solve the issue of unobservable confounds, but remain susceptible to this attribution problem. (Note that in this research the measured outcome is affect toward an institution.)

Given these challenges, an alternative design would randomize some treatment thought to proxy a legitimating action by an authority. This “naive” randomization approach, depicted in Fig. 1B (with the random variable $Z$ determining treatment assignment), resolves the problem of omitted variables bias and can recover a valid estimates of the total average treatment effect of $A$ on $C$. But absent further assumptions it cannot be used to distinguish the causal mechanism underling this effect, as it may arise either through the material incentives channel ($A$ to $M$ to $C$) or via a legitimacy mechanism ($A$ to $C$). Alternatively, one could restrict attention to treatments that have no effect on the citizens’ material incentives, eliminating the $M$ node altogether, but this introduces two problems of ecological validity. First, it precludes studying any legitimating action related to the fairness of payoff-relevant factors in a citizens’ compliance choice. Insofar as fairness is nearly always defined with respect to materially relevant outcomes such as accuracy or bias in the application of punishment, this domain restriction is quite severe. Second, randomization undermines the ability of citizens to form beliefs about the authority herself – beliefs that surely form the foundation of any account of her legitimacy. In the extreme, for example, a person compelled to take an action does not signal anything about herself by that choice.

Our main experimental approach to circumventing these challenges draws on recent advances in design-based approaches to identifying causal mediation effects (22, 23) and is depicted in Fig. 1C. As with the naive randomization approach, we circumvent problems of omit-
ted variables bias by randomly matching subjects assigned the citizen and authority roles and taking advantage of authorities’ choices of whether to seek more or less legitimating actions to induce variation in citizens’ experiences. Second, we sever the deterministic link between the authority’s legitimating action and the citizen’s material incentives by introducing a stochastic mediating variable \((Z)\) triggered by the authority’s action, which in turn determines the level of the material incentives \(M\).

Suppose that \(Z\) is more likely to take on values that lead to high levels of \(M\) when \(A = 1\) than when \(A = 0\). (Our main experiment is slightly more involved, but the intuition carries over.) If we condition on a value of \(Z = z\) that is feasible both when \(A = 1\) and \(A = 0\), comparing values of \(C\) under those different values of \(A\) identifies the direct effect of \(A\) on \(C\) given \(Z = z\). Conditioning on \(A\) and comparing values of \(C\) under different levels of \(M\), by contrast, identifies the conditional causal mediation effect of material incentives on citizen behavior in either the presence \((A = 1)\) or absence \((A = 0)\) of the authority’s action.

Informally, then, this design permits us to identify the effect of an authority’s “good faith effort” to improve the fairness of her enforcement work on citizens’ subsequent compliance decisions. Further, the design permits a comparison of the magnitude of this “legitimating” effect to the effect of changes in material incentives on compliance decisions. That being said, the design as stated is limited in the extent to which it permits us to decompose the various psychological mechanisms that may contribute to the legitimating effect – an issue that animates our second experiment, described in greater detail below.

The Intrinsic Channel: Possible Mechanisms. While the identification strategy described in the preceding section permits us to distinguish between the extrinsic and intrinsic channels, it does not, by itself, distinguish among the various psychological mechanisms that might animate the latter. Here, we consider three. (A fourth, which involves the authority’s action as a coordi-
The first, normative legitimation, comes closest to the idea of legitimate authority described in the social scientific literature cited above as well as by normative political theorists (25–27). Underpinning this account is the notion that an authority’s legitimacy depends on the view of its subjects that it is worthy of recognition as such, and that this recognition entails some sense of obligation to comply with the authority’s edicts. One source of this perception on the part of subjects is the apparent procedural fairness of the authority, which can arise from a number of factors, including accuracy, correctability, consistency, and unbiasedness (10). Behaviorally, the normative legitimation mechanism would involve citizens updating their beliefs about the quality of the authority herself (perhaps based on a costly action taken by that authority), with the resultant sense of esteem fostering an intrinsic motivation to comply.

According to a second, conditional cooperation mechanism, the authority’s behavior conveys information to citizens about the prevalence of pro-social types among other citizens, with greater regard for those citizens inducing stronger motivation to contribute out of an anticipation that those citizens will, too, contribute (28). This mechanism is plausible in our experimental setting because authorities and citizens are drawn from the same population of subjects. It also implies several ancillary empirical implications, which we test below.

A third mechanism involves direct reciprocity to the authority. By choosing a costly action that improves the fairness of enforcement, the authority might be seen as having done right by citizens. Citizens, may in turn feel intrinsically motivated to “pay back” the authority for having undertaken this investment (29). In our first experiment, citizens could achieve this goal by contributing to the public good, because the authority herself benefits from these contributions. Our second experiment is explicitly designed to distinguish the direct reciprocity mechanism.

We do not mean to suggest that this constitutes an exhaustive accounting of all conceivable
mechanisms; and there is also little reason to believe ex ante that only one mechanism is relevant on the broad topic of the relationship between authorities and citizens. That said, our analysis below provides an approach to adjudicating among these mechanisms, while also suggesting how one might go about testing, in future work, additional ones.

**Studying Legitimacy Using Behavioral Experiments**

In keeping with prior experimental research on compliance with authority, subjects in both experiments participated in a linear public goods contribution game with a centralized enforcer. The Methods section describes the experimental protocol in detail, including the payoff functions and information structure. A brief summary follows here.

Contribution is individually costly for citizens, but citizens are collectively better off if all players contribute than if none do. This is analogous to a situation in which people would individually improve their well-being by breaking a law or social norm (in the absence of punishment), but have higher welfare when everyone complies with the law than when no one does.

The authority chooses an enforcement rule and a level of investment in the accuracy of information available to her. “Accuracy” simply refers to the probability the enforcer observes a signal that a given citizen kept (contributed) their tokens when they in fact kept (contributed) them. In keeping with the discussion of our identification strategy above, this investment generates superior information for the authority only probabilistically; it is this feature that allows us to decouple subjects’ intrinsic and extrinsic incentives.

At the beginning of each period, subjects were each given an endowment of tokens and randomly assigned to a group of five people, of which four were randomly assigned as citizens (“Role A”) and one as an authority (“Role B”). Each period of Experiment 1 consisted of one play of the following extensive form game:

1. Authority chooses “big” or “small” investment in accuracy.
2. Accuracy level given authority’s investment determined via randomization.

- If the authority makes the big investment in accuracy, realized accuracy is equally likely to be “high” or “medium.”
- If the authority makes the small investment in accuracy, realized accuracy is equally likely to be “low” or “medium.”

3. Authority learns realized accuracy level and chooses one of four possible enforcement rules. We will focus our attention on the large majority of cases in which the authority chose the intuitive “punish according to signal” (PATS) rule, which deducts tokens from each citizen who appeared not to contribute to the public good (based on a potentially erroneous signal).

4. Each citizen learns the authority’s investment, enforcement rule, and realized accuracy level, and chooses whether or not to contribute their endowment to the public good. (A valuable extension, which we leave to future work, would allow for intermediate contributions.)

5. Signals are generated based on citizens’ choices; authority’s chosen enforcement rule is automatically implemented; citizens and authority each receive payoffs.

Experiment 2 is identical to Experiment 1 with two changes. First, between stages 3 and 4 of the timeline above, there is an additional randomization: half the time, the authority’s compensation is 40% of the value of the common pot, as in Experiment 1; the other half the time it is a flat fee of 20 tokens. The second randomization permits us to directly assess the plausibility of the reciprocity mechanism. Fig. 1D depicts the underlying logic of this design, with \( R \) denoting reciprocity motivations.
The randomization of compensation, denoted $Z'$ in the figure, takes place after the authority has made 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 know both 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.

Experiment 2 differs from Experiment 1 in one further respect. We improve the statistical power of the experiment by including only two (rather than three) levels of accuracy: High and Low. A big investment by the authority makes high accuracy more likely.

**Results**

**Results for Experiment 1**

**Authority Behavior.** The data consist of 360 group-period interactions. Table 1 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 sig-
nal) 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%). (Two players accounted for 47% of the cases of anti-PATS, while eight players chose the anti-PATS rule only once.) 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 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 (30), contributions do not diminish over time. Fig. 2 displays data on group-level contributions by period (tokens 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 the others (Never Punish, Always Punish, and anti-PATS) it is, 1.0, 1.4, and 0.8 respectively. (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 the other enforcement rules may emerge because of large differences in the 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 and contributors were punished 42% of the time.

**Effect of the Legitimating Action.** At the moment citizens are choosing whether 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. (Choices of other citizens drop out of the material payoffs calculation.) Additionally, they have also observed whether the authority made the big investment in accuracy. Conditional on the realized accuracy level, this choice by the authority is materially irrelevant. Table 2 describes the conditions under which citizens make their contribution choices under the PATS rule. (The notation \( C_{r,k} \) denotes sample average group-level contribution levels in row \( r \) [investment] and column \( k \) [accuracy].) The comparison of greatest interest is

\[
C_{B,M} - C_{S,M}.
\]

Given the PATS enforcement rule, this holds all material motivations to contribute constant, thus identifying the causal effect of an authority’s legitimating action.

The relevant data are displayed graphically in Fig. 3, which plots group average contributions by the authority’s investment and realized accuracy level, conditional on the PATS enforcement rule. The effect of the authority’s legitimating action is calculated by comparing the dark and light gray bars (bar #3 and bar #2, respectively) in the figure. 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 corresponds to a 14% increase in the contribution rate.

**Extrinsic Accuracy Effects.** We can also make two comparisons that isolate the effect of improvements in accuracy via the extrinsic channel. The first, using the notation of Table 2, is $C_{S,M} - C_{S,L}$. This identifies the effect of a change from Low to Medium accuracy, holding the authority’s investment choice constant at Small. This is calculated by comparing the pale gray column (bar #2, medium accuracy, small investment) in Fig. 3 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 statistically significant at conventional levels and implies that the change in accuracy from low to medium induces a 60% increase in contributions.

The second comparison is $C_{B,H} - C_{B,M}$, which identifies the effect of a change from Medium to High accuracy, holding the authority’s choice constant at Big. This 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%.

Comparing the effect of the legitimating action with these extrinsic accuracy effects, it appears that the former is comparable in magnitude to the extrinsic accuracy effect conditional on the large investment, but smaller than the extrinsic effect conditional on the small investment. We caution against making too much of this comparison, as the relative effects of extrinsic and intrinsic motivations will themselves vary depending on the parameters of the experiment or aspects of real-world contexts.
Robustness. Other factors in the experiment beyond those described above may 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. (Even if players condition their behavior on their prior group-level experiences, groups are reshuffled across rounds so those expectations should be identical across groups in subsequent rounds.)

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

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

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
the big investment in this specification is around 10 percentage points ($p < 0.01$, two-tailed).

(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.) 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 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, to ensure that our results are not driven by a systematic misunderstanding of the experimental protocol.

Our most conservative analysis appears in columns (7) and (8). In these specifications, we include individual-level fixed effects. 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 con-
tribute, we continue to find evidence that the authority’s legitimating choice directly increased the likelihood a player contributed to the public good.

Assessing the Conditional Cooperation Mechanism.  Experiment 2 below is designed specifically to test for the presence of the direct reciprocity mechanism. Ancillary results from Experiment 1 permit us to assess the operation of the conditional cooperation mechanism as well. Specifically, if this mechanism were the primary driver of our Experiment 1 results, we would anticipate, first, that conditioning on the citizen’s prior experience with group contributions and authority behavior would attenuate the estimate of the intrinsic effect. The data do not bear out this hypothesis, however. Specifically, columns (3) and (4) of Table 3 include direct measures of each citizen’s prior experience, a proxy for each player’s beliefs about the behavior of others. Doing so does not reduce the estimated effect of interest.

Additionally, although the authority’s choice may provide novel information about other citizen’s proclivities, we would expect it to be most 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 our estimate for the intrinsic effect is robust and stable across subsets of the periods of the game. Fig. 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.

The persistence of this effect over time is also reassuring because it mitigates against reciprocity-based explanations. 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.

More formally, while we lack statistical power to conduct period-specific comparisons, par-
titioning 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). This evidence does not prove that conditional cooperation is unimportant. However, the relative stability in the effect of the legitimating action over time strongly suggests that this mechanism is not the primary driving force behind our key result from Experiment 1.

Results for Experiment 2

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 stakes of the choice in this setting. Additionally, fewer subjects in the role of the authority chose the big investment: 33% vs. 60%. When authorities chose the PATS enforcement rule, they chose the big investment 43% of the time (compared to 69% of the time in Experiment 1). 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 Fig. 5).

Citizen Behavior Aggregated Across Authority Compensation Mechanisms. With only two realized levels of accuracy that may be achieved under both the big and small authority investment, we can estimate the effects of the legitimating action, for each compensation mechanism, both under Low accuracy ($C^m_{B,L} - C^m_{S,L}$), modifying the notation from above with the superscript $m \in \{\text{flat, benefit}\}$ and High accuracy ($C^m_{B,H} - C^m_{S,H}$).

Before disaggregating citizen behavior by the authority’s compensation mechanism, we present summary data suggesting that the findings of the first experiment replicate in this al-
ternative environment. Fig. 6 displays average group contribution rates conditional on the PATS enforcement rule, for different authority investments and realized levels of accuracy.

First, we consider the effect of the legitimating action, calculated at the group-period level and aggregating across authority compensation mechanisms. The effect conditional on high accuracy is obtained by comparing the black and dark gray bars (bar #4 and bar #3, respectively) in Fig. 6. We find the effect to be a statistically insignificant difference of 0.12 contributions ($p = 0.40$, two-tailed). This insignificant result is not especially surprising: conditional on the high level of accuracy, subjects should strictly prefer to contribute to the public good based on extrinsic motivations alone; the data confirm this hypothesis.

More relevant for the current discussion is the effect of authority investment conditional on the low accuracy level (comparing bar #2 and bar #1). Here, we observe a statistically significant 27% increase in contributions, from an average of 1.74 to 2.2.

The extrinsic accuracy 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 statistically significant at conventional levels: conditional on the small investment, an increase from low to high accuracy nearly doubles the contribution rate, from 1.74 to 3.35. The extrinsic accuracy effect conditional on the big investment is smaller in magnitude, but still statistically significant: an increase from 2.20 to 3.47, or 58%.

**Effects of the Legitimating Action 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 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 effect of the legitimating action 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 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.

**Discussion**

The question of what makes authority legitimate has occupied empirical social scientists since Weber’s (31) seminal writings on the subject in the early 20th century. Recently, the prominent
behavioral research on the subject has emerged at the intersection of social psychology (17,32) and criminal justice (33), with particular attention paid to the relationship between perceptions of procedural justice (34) and legitimacy. Political science research on legitimate authority has paid closest attention to attitudes on legitimacy and the closely related notion of “diffuse support” (35), focusing on support for courts (36,37) and international organizations (38). (Note that the question of legitimacy is related to, but analytically distinct from, institutional trust, as an individual might trust an institution owing to its legitimacy or an assessment of its material performance.) A common thread in these two political science research traditions is the question of how actors with limited enforcement power achieve deference in spite of their politically tenuous positions.

This paper makes several contributions to the study of legitimate authority and its relationship with subordinate behavior. Foremost, we demonstrate that subordinates can be motivated to comply with an authority as a consequence of changes to her perceived legitimacy, holding constant purely instrumental motivations. This phenomenon contributes fundamentally to our understanding of how perceptions of fairness affect behavior, a dynamic that persistently plays out in salient policy domains such as policing.

Methodologically, the contribution of this paper is to formalize the challenge of separately identifying legitimating effects from material motivations for compliance, and to propose an experimental protocol that solves this attribution problem. The innovation of our experimental designs is to isolate the legitimating effect of an authority’s effort to secure a more procedurally fair institution from the effects of the institution 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 that an authority’s legitimating action substantially increases the willingness of citizens to contribute to a collective good. This effect is robust to different sample
restrictions and statistical modeling approaches. 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.

This is, to our knowledge, the first causally credible evidence about the effects of an authority’s legitimating policy choices on citizen behavior. Unlike work that seeks to measure perceptions of legitimacy and other factors as predictors of reported behavior and attitudes, we exogenously manipulate perceptions of legitimacy holding fixed the material incentives created by an authority’s policy choice. Thus, we need not rely on “as-if” randomness in perceptions of legitimacy, with the attendant risks of reverse causality and omitted variables bias, to estimate the importance of legitimacy. Simultaneously, we are also freed from concerns about measurement error and model specification that arise when using measures of non-legitimacy factors to “control for” those factors so as to isolate legitimacy effects. Instead, our design-based approach provides clear evidence about the direct importance of legitimacy in shaping citizen behavior. Despite the strong prior beliefs that many people hold about the importance of legitimacy, the empirical evidence to date for its importance is not dispositive.

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. While the incentives deployed in our analysis are modest, survey responses are entirely costless, and as such analysis relying
solely on those measures may be expressive rather than predictive of actual behaviors (in addition to raising problems of causal inference and measurement error). It is therefore notable that almost all extant evidence about the importance of legitimacy is survey-based.

That being said, a clear benefit of survey-based approaches is their scalability to broader populations that are more representative of the citizenry at large. With this in mind, a natural future direction for this research would involve either the use of incentivized surveys or lab-in-the-field experiments. Researchers may also be able to develop surveys employing benchmarks of “rational” behavior in ways typically missing from earlier studies.

Third, our specific experimental design also allows us to make progress in understanding the psychological origins of legitimacy. We show that authorities who attempt to obtain more accurate institutions are more legitimate in the eyes of their citizens. Of course there are numerous potential sources of improved legitimacy aside from accuracy. We focus here on those choices that are under an authority’s control. For example, the authority may or may not take actions that appear biased against a member of a specific group. The experimental approach presented here is sufficiently flexible to study other sources of legitimacy, a task for future research.

The stylized environment in which we are able to implement this design may depart from non-experimental settings in a number of respects. Two features of the authority essential to the design that deserve comment are her ability to pre-commit to an enforcement rule and her unlimited enforcement capacity. In the absence of pre-commitment, citizens would be able to make inferences about likely enforcement strategies from the (ex ante) procedural investment, further entangling instrumental and legitimacy-based mechanisms. Likewise, if an authority’s capacity is limited, in many circumstances there will be multiple equilibria in which expectations about others’ behavior come into play. In such an environment, authorities’ legitimating actions could serve an equilibrium selection role, affecting citizen behavior by coordinating expectations about others’ behavior.
Additionally, the data we present here draws from selected samples interacting in a controlled and stylized environment. A key task for future research is to understand whether there is heterogeneity in behavioral responses to an authority’s choices that differs between these on campus samples and those drawn from the population at large. We note that evidence that experimental estimates are dramatically affected by sample composition is limited, but believe it would be useful to have more broadly representative estimates of average treatment effects. In terms of the stylized environment used in our experiments, the key advantage of this setup is that this control is essential to make causal claims. While it is tempting to add additional features that capture the complexity of “real world” citizen-authority interactions, doing so would imperil our ability to make progress in understanding the causal role of legitimacy. Indeed, it is precisely because the wealth of evidence drawn from observations of actual citizen-authority interactions is compatible with multiple theoretical explanations that more credible evidence is needed. Skeptics of legitimacy-based accounts (39) have pointed to this empirical ambiguity as a reason to discard legitimacy-based explanations altogether. In this case, while abstraction removes some of the texture and complexity of real life, the advantages in terms of theory testing are real and substantial.

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.

**Methods**

**Experimental Protocol: Additional Details**

**Accuracy.** In Experiment 1, depending on its authority’s choice, each group was 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.

In Experiment 2, there are two accuracy levels: High (20% error rate) and Low (40% error rate). Conditional on the small investment, accuracy is Low with 75% probability and High with 25% probability; conditional on the big investment, those probabilities were reversed. (Note that these are not error rates drawn from real-world examples, but instead designed to create known benchmark behavior given the punishment regime in place in the experimental
setting – see below.)

**Enforcement Rule.** In both experiments, 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

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. Pre-commitment 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 Experiment 1 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. 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 our identification goals. Very strong or very weak material incentives to contribute would, by contrast, obscure the possibility that legitimating actions per se might have a causal effect on citizen behavior.

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. (In the presence of idiosyncratic shocks to citizens’ motivations to contribute, PATS is the only enforcement rule consistent with equilibrium play.) 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 on cases in which the authority chooses PATS (and thus holding constant the enforcement rule).

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).

In Experiment 2, there is no realized level of accuracy that would make a subject motivated purely by extrinsic incentives completely 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.

**Implementation Details**

In both experiments, subjects interacted anonymously via networked computers. The experiments were programmed and conducted using the software z-Tree (40) and the experimental protocols were approved for compliance with Human Subjects regulations by the relevant Institutional Review Boards. 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 (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.

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.

For Experiment 1, we conducted four experimental sessions at the NYU Center for Experimental Social Science Lab and two sessions at the Yale Behavioral Research lab. 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, 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 $2.50), with a maximum of $30.07 and a minimum of $17.67. 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 27% receiving a perfect score.

For Experiment 2, we collected data on 110 subjects gathered during six sessions conducted at the NYU site and one at the Yale lab. 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.

References


Acknowledgments

The authors declare no competing interests. The authors all participated in research and experimental design, data collection, data analysis, and writing. We thank 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. This material is based upon work supported by the National Science Foundation under Grant Nos. 1555494 and 1556047. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Additional material necessary to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. A complete replication archive including source data and statistical code necessary to recreate our analysis is available at https://doi.org/10.7910/DVN/SPIMAV.
### Tables

<table>
<thead>
<tr>
<th>Enforcement Rule</th>
<th>PATS</th>
<th>Anti-PATS</th>
<th>Never Punish</th>
<th>Always Punish</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Investment</td>
<td>81</td>
<td>7</td>
<td>40</td>
<td>18</td>
<td>146</td>
</tr>
<tr>
<td>Big Investment</td>
<td>185</td>
<td>8</td>
<td>15</td>
<td>6</td>
<td>214</td>
</tr>
<tr>
<td>Total</td>
<td>266</td>
<td>15</td>
<td>55</td>
<td>24</td>
<td>360</td>
</tr>
</tbody>
</table>

**Table 1. Authority Choices in the Baseline Experiment: Group-Level Data.** PATS enforcement rule: deduct tokens from each citizen for whom a signal of “kept” was generated (see text). Anti-PATS rule: deduct tokens from each citizen for whom a signal of “contributed” was generated. Always and Never Punish rules are self-explanatory.
<table>
<thead>
<tr>
<th>Authority Investment</th>
<th>Realized Accuracy Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Small</td>
<td>$\overline{C}_{S,L}$</td>
</tr>
<tr>
<td>Big</td>
<td>$-$</td>
</tr>
</tbody>
</table>

**Table 2. Identification of Legitimacy Effect, Holding Punishment Strategy Constant.** Cell entries denote sample average group contribution rates.
<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>Authority Investment</td>
<td>0.101</td>
<td>0.109</td>
<td>0.115</td>
<td>0.125</td>
<td>0.118</td>
<td>0.111</td>
<td>0.104</td>
<td>0.124</td>
</tr>
<tr>
<td>(1=big, 0=small)</td>
<td>[0.038]</td>
<td>[0.037]</td>
<td>[0.037]</td>
<td>[0.038]</td>
<td>[0.048]</td>
<td>[0.052]</td>
<td>[0.039]</td>
<td>[0.046]</td>
</tr>
<tr>
<td>Average group contributions</td>
<td>0.072</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.078</td>
</tr>
<tr>
<td>prior to this period (0-4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[0.030]</td>
</tr>
<tr>
<td>Average investment experience</td>
<td>-0.083</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.047</td>
</tr>
<tr>
<td>prior to this period (0-1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[0.081]</td>
</tr>
<tr>
<td>Constant</td>
<td>0.731</td>
<td>0.635</td>
<td>0.588</td>
<td>0.768</td>
<td>0.929</td>
<td>0.770</td>
<td>0.747</td>
<td>0.597</td>
</tr>
<tr>
<td></td>
<td>[0.032]</td>
<td>[0.063]</td>
<td>[0.105]</td>
<td>[0.098]</td>
<td>[0.039]</td>
<td>[0.063]</td>
<td>[0.025]</td>
<td>[0.139]</td>
</tr>
<tr>
<td>Observations</td>
<td>560</td>
<td>560</td>
<td>540</td>
<td>540</td>
<td>404</td>
<td>338</td>
<td>447</td>
<td>428</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.01</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
<td>0.08</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Number of unique subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Period fixed effects</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Subject fixed effects</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Table 3. Estimated Effect of Authority’s Investment Choice, Given Medium Accuracy and PATS Enforcement Rule. 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.
Table 4. **Authority Choices in the Second Experiment: Group-Level Data.** See notes to Table 1 for explanation of terms.

<table>
<thead>
<tr>
<th>Enforcement Rule</th>
<th>PATS</th>
<th>Anti-PATS</th>
<th>Never Punish</th>
<th>Always Punish</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Investment</td>
<td>142</td>
<td>35</td>
<td>82</td>
<td>35</td>
<td>294</td>
</tr>
<tr>
<td>Big Investment</td>
<td>108</td>
<td>5</td>
<td>13</td>
<td>20</td>
<td>146</td>
</tr>
<tr>
<td>Total</td>
<td>250</td>
<td>40</td>
<td>95</td>
<td>55</td>
<td>440</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.370</td>
<td>0.403</td>
<td>0.369</td>
<td>0.403</td>
<td></td>
</tr>
<tr>
<td>(1=high, 0=low)</td>
<td>[0.032]***</td>
<td>[0.027]***</td>
<td>[0.032]***</td>
<td>[0.027]***</td>
<td></td>
</tr>
<tr>
<td>Authority Investment</td>
<td>0.066</td>
<td>0.063</td>
<td>0.076</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>(1=high, 0=low)</td>
<td>[0.031]**</td>
<td>[0.028]**</td>
<td>[0.041]*</td>
<td>[0.035]</td>
<td></td>
</tr>
<tr>
<td>Average group contributions</td>
<td></td>
<td>0.154</td>
<td></td>
<td>0.154</td>
<td></td>
</tr>
<tr>
<td>in prior periods (0-4)</td>
<td></td>
<td>[0.023]***</td>
<td></td>
<td>[0.023]***</td>
<td></td>
</tr>
<tr>
<td>Average investments</td>
<td>0.001</td>
<td></td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in prior periods (0-1)</td>
<td></td>
<td>[0.078]</td>
<td></td>
<td>[0.078]</td>
<td></td>
</tr>
<tr>
<td>Authority benefits × Investment</td>
<td></td>
<td>-0.017</td>
<td>-0.026</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.042]</td>
<td>[0.036]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authority benefits</td>
<td></td>
<td>-0.019</td>
<td>0.013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.444</td>
<td>0.102</td>
<td>0.452</td>
<td>0.113</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.025]***</td>
<td>[0.072]</td>
<td>[0.031]***</td>
<td>[0.074]</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1,000</td>
<td>960</td>
<td>1,000</td>
<td>960</td>
<td></td>
</tr>
<tr>
<td>R-squared</td>
<td>0.181</td>
<td>0.245</td>
<td>0.182</td>
<td>0.246</td>
<td></td>
</tr>
<tr>
<td>Period fixed effects</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>

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

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.
Figures

Fig. 1. Directed Acyclic Graphs Demonstrating Challenges to Identifying Legitimating Effects of Authority Choices, and Potential Solutions. Panel (1) depicts two challenges with uncovering the legitimating effects of an authority’s actions. Panel (2) depicts an experimental approach that randomizes $A$. Panel (3) depicts our stochastic mediation approach (Experiment 1) to the attribution problem. Panel (4) depicts an augmented approach (Experiment 2) to isolate the legitimation mechanism. $A$ denotes authority’s action; $M$ material incentives; $C$ citizen’s compliance choice; $U$ unobserved confounders; $R$ the authority’s reciprocity motivations; and $Z$ and $Z'$ random assignment variables.

Fig. 2. Group Contributions by Period. Note: Data vertically jittered to enhance clarity of presentation. Dashed line is local polynomial smoother, with the gray shaded area denoting the associated 95% confidence interval.

Fig. 3. Group-level Contributions by Investment Decision and Accuracy Levels Given the Authority Chooses Punish According to Signal, Experiment 1. Note: Averages with bootstrapped 95% confidence intervals. (Non-overlapping confidence intervals are sufficient, but not necessary, for statistical significance of reported differences.)

Fig. 4. Effect of the Legitimating Action: Persistence Over Time. Data jittered to enhance clarity of presentation. Observations are group contributions given medium accuracy and PATS enforcement strategy. Lines are local polynomial smoothers.

Fig. 5. Group Contributions by Period, Experiment 2. Data vertically jittered to enhance clarity of presentation. Dashed line is local polynomial smoother, with the gray shaded area
denoting the associated 95% confidence interval.

**Fig. 6. Group-level Contributions by Investment Decision and Accuracy Levels Given the Authority Chooses Punish According to Signal, Experiment 2.** Averages with bootstrapped 95% confidence intervals. (Non-overlapping confidence intervals are sufficient, but not necessary, for statistical significance of reported differences.)
A. Two Challenges to Identification

B. Naive Randomization Approach

C. Stochastic Mediator

D. Dual Randomization
Average Group Contribution

1. Low Accuracy
   Small Investment

2. Medium Accuracy
   Small Investment

3. Medium Accuracy
   Big Investment

4. High Accuracy
   Big Investment

- diff. = 1.096
  p < 0.001

- diff. = 0.405
  p = 0.008

- diff. = 0.341
  p < 0.001
Contributions

Period

Big Investment

Small Investment
1. Low Accuracy
   Small Investment
2. Low Accuracy
   Big Investment
3. High Accuracy
   Small Investment
4. High Accuracy
   Big Investment

diff. = 0.46
p = 0.049

diff. = 0.12
p = 0.39
For Online Publication:
Appendix: Instructions for Baseline Experiment

Instructions

This is an experiment on decision making. In the following experiment you will make a series of choices. At the end of the experiment, you will be paid depending on the specific choices that you made during the experiment and the choices made by other people. All of your interactions will be through the computer terminals at which you are sitting, and your true identity will never be revealed to any other person in the laboratory. If you follow the instructions and make appropriate decisions, you may make an appreciable amount of money. Please remain silent and listen carefully to the instructions.

In addition to the show-up fee of $7, during the course of the experiment you will have the opportunity to earn “tokens” that will be converted into dollars at the end of the experiment. The conversion rate is:

30 tokens = 1 dollar

The experiment consists of 20 periods, all of which have the same structure.

At the beginning of each period:
- You will be randomly assigned into a new group of five people
- You will be randomly assigned to one of two roles:
  - Role A. Within a given group, four people are assigned to Role A. These four people are each randomly assigned a temporary ID number -- 1, 2, 3, or 4. Each person in Role A makes a simple choice: whether to allocate his or her individual supply of tokens to a common pot, or to keep these tokens for him- or herself.
  - Role B. Within a given group, one person is assigned to Role B. The person in Role B will choose a rule that may result in some people in Role A having their payoffs reduced. The person in Role B will also choose either a “Small Investment” or a “Big Investment” in receiving accurate information about the choices of people in Role A.

Specifically, each of the 20 periods consists of three separate stages, which always proceed in the following order:
1. The person in Role B makes his or her choices about the level of investment in accurate information and the payoff reduction rule.
2. Each person in Role A chooses whether to allocate an individual supply of tokens to a common pot, or to keep these tokens for him- or herself.
3. Some people in Role A may have their payoffs reduced, depending in part on the choices already made by the person in Role B in Stage 1.

A more complete description of this process now follows.
In Stage 1, the person in Role B makes choices about (1) investing in accurate information, and (2) a rule that potentially reduces the payoffs of some people in Role A.

(1) Investing in Accurate Information

As noted above, each person in Role A will eventually make a simple choice: whether to allocate an individual supply of tokens to a common pot, or to keep these tokens for him-or herself. After these decisions are made, the person in Role B will receive some information about the decision of each person in Role A, listed by his or her temporary ID number (1, 2, 3, or 4). This information may be “Low Accuracy,” “Medium Accuracy,” or “High Accuracy”:

- “Low Accuracy” Information has a “40\% error rate”
- “Medium Accuracy” Information has a “25\% error rate”
- “High Accuracy” Information has a “10\% error rate”

An error rate describes how likely it is that the information received about a given person in Role A is inaccurate. As an example, suppose that the error rate is 25\%. In this case, if a given person in Role A (say, the person with ID number 2) in fact *allocated* his or her tokens to the common pot, there is a 25\% likelihood that the person in Role B will receive an *incorrect* message that this person “kept” his or her tokens, but a 75\% likelihood that the person in Role B will receive a *correct* message that this person “allocated” his or her tokens. Similarly, if a given person in Role A (say, the person with ID number 2) in fact *kept* his or her tokens, there is a 25\% likelihood that the person in Role B will receive an *incorrect* message that this person “allocated” his or her tokens, but a 75\% likelihood that the person in Role B will receive a *correct* message that this person “kept” his or her tokens.

Whether the information received is “Low Accuracy,” “Medium Accuracy,” or “High Accuracy” will depend on an investment decision made by the person in Role B in Stage 1. Specifically, the person in Role B must choose either a “Small Investment” or a “Big Investment”.

If the person in Role B chooses a “Small Investment”:

- He or she pays no cost.
- He or she receives “Low Accuracy” information with 50\% likelihood, but receives “Medium Accuracy” information with 50\% likelihood.

If the person in Role B chooses a “Big Investment”:

- He or she pays a cost of 4 tokens.
- He or she receives “High Accuracy” information with 50\% likelihood, but receives “Medium Accuracy” information with 50\% likelihood.

The person in Role B will be informed whether his or her investment choice resulted in “Low Accuracy,” “Medium Accuracy,” or “High Accuracy” information before choosing a payoff-reduction rule.
(2) Payoff-Reduction Rule

In Stage 1, the person in Role B also chooses a rule that potentially reduces the payoffs of some people in Role A. As noted above, the person in Role B will ultimately receive some information about the decision of each person in Role A, which may or may not be accurate. The rule selected by the person in Role B will eventually use this information.

Specifically, the person in Role B will choose one of the four following payoff-reduction rules:

- (1) Reduce by 24 tokens the payoffs of every person in Role A, regardless of the message received about that person.
- (2) Reduce by 24 tokens the payoffs of those people in Role A about whom a message of “kept” is received, and do not reduce the payoffs of those people in Role A about whom a message of “allocated” is received.
- (3) Do not reduce the payoffs of those people in Role A about whom a message of “kept” is received, and reduce by 24 tokens the payoffs of those people in Role A about whom a message of “allocated” is received.
- (4) Do not reduce the payoffs of any person in Role A, regardless of the message received about that person.

Remember that the message received about each person in Role A may or may not be accurate, as described above.

Summary

In Stage 1, the person in Role B chooses either a “Small Investment” or a “Big Investment.” The person in Role B also chooses one of the four payoff-reduction rules listed above.

Once Stage 1 is complete, the people in Role A learn the following outcomes of Stage 1:

- Whether the person in Role B chose a “Small Investment” or a “Big Investment”
- Whether this investment resulted in “Low Accuracy” Information, “Medium Accuracy” Information, or “High Accuracy” Information
- Which payoff-reduction rule was chosen by the person in Role B

Stage 2

In Stage 2, each person in Role A receives 20 tokens. Each person in Role A must then decide whether to allocate these 20 tokens to a common pot, or to keep them for him- or herself. Each person in Role A must choose either to allocate all 20 of the tokens or keep all 20 of the tokens.

The payoff to a person in Role A in Stage 2 is composed of two parts:

- The number of tokens the person keeps for him- or herself
  PLUS
• 0.4 times the number of tokens that all people in Role A allocate to the common pot (including tokens the person allocates him- or herself).

That is, the payoff to a person in Role A in stage 2 can be written as

\[ \text{Stage 2 payoff to person in Role A} = \text{(tokens kept)} + 0.4 \times \text{(total tokens allocated to common pot by people in Role A)}. \]

Although the person in Role B does not make a choice in Stage 2, he or she also receives Stage 2 payoffs that depend on the choices made by people in Role A. These payoffs are composed of two parts:

• An automatic supply of 20 tokens
• 0.4 times the number of tokens that all people in Role A allocate to the common pot.

\[ \text{Stage 2 payoff to person in Role B} = 20 + 0.4 \times \text{(total tokens allocated to common pot by people in Role A)}. \]

Therefore, every token kept by a given person in Role A increases that person’s Stage 2 payoffs by one token (and does not contribute to the payoffs of other group members). Every token allocated to the common pot increases the Stage 2 payoffs of every person in Role A by 0.4 tokens, and also increases the Stage 2 payoffs of the person in Role B by 0.4 tokens.

Consider the following examples:

• Suppose that every person in Role A keeps all of his or her 20 tokens for himself or herself. Then the Stage 2 payoffs of each person in Role A will be equal to 20 + (0.4 \times 0) = 20 tokens. The Stage 2 payoffs of the person in Role B will be equal to 20 + (0.4 \times 0) = 20 tokens.

• Suppose that every person in Role A allocates all of his or her 20 tokens to the common pot. Then the Stage 2 payoffs of each person in Role A will be equal to 0 + (0.4 \times 80) = 32 tokens. The Stage 2 payoffs of the person in Role B will be equal to 20 + (0.4 \times 80) = 52 tokens.

• Suppose that two people in Role A each allocate their 20 tokens to the common pot, and two people in Role A each keep their 20 tokens for themselves. In total, 40 tokens are allocated to the common pot by the people in Role A as a whole. Then each person in Role A receives 0.4 \times 40 = 16 tokens from the common pot. The two people in Role A who each allocated 20 tokens to the common pot (keeping none for themselves) would therefore have Stage 2 payoffs equal to 0 + (0.4 \times 40) = 16 tokens. The two people in Role A who each allocated no tokens to the common pot (keeping 20 for themselves) would therefore have Stage 2 payoffs equal to 20 + (0.4 \times 40) =
36 tokens. The person in Role B would receive Stage 2 payoffs equal to $20 + (0.4 \times 40) = 36$ tokens.

**Stage 3**

Neither the person in Role B nor the people in Role A make further choices in Stage 3. Messages about whether each person in Role A “allocated” or “kept” his or her tokens in Stage 2 are generated as described above, taking into account whether there is “Low Accuracy” Information, “Medium Accuracy” information, or “High Accuracy” Information.

The payoff-reduction rule chosen by the person in Role B in Stage 1 then takes effect, using the messages to the person in Role B about each person in Role A. Once Stage 3 is complete, the people in Role A and the person in Role B all learn the following outcomes of Stages 2 and 3:

- Whether each person in Role A in fact “allocated” or “kept” his or her tokens
- The message generated about each person in Role A, either “allocated” or “kept”
- The amount, if any, by which the payoffs of each person in Role A were reduced.

**Summary of Net Payoffs for a Period**

For a person in Role A, the following is calculated:

- Stage 2 payoffs (from common pot and tokens kept)….
- …MINUS the number of tokens, if any, reduced by the payoff-reduction rule in Stage 3

The result is the net payoffs for the period for a person in Role A.

For the person in Role B, the following is calculated:

- Stage 2 payoffs (from common pot and automatic token supply)….
- …MINUS the number of tokens, if any, spent in Stage 1.

The result is the net payoffs for the period for the person in Role B.

Regardless of your Role, you will see your net payoffs for the period on your screen once that period is complete.

**Conclusion**

This concludes the description of the choices that are made and the payoffs that are earned in one period. Remember that you will be randomly assigned into a new group, and a new role, at the beginning of each period. This process will be repeated until all of the 20 identical periods are completed. Your payment for the experiment will consist of the sum of your payoffs from all 20 periods, plus the show-up fee.

We ask everyone to remain silent until the end of the last period and then to await further instructions. If you have any questions, please raise your hand while remaining silent.
Appendix: Experiment #1 Comprehension Quiz Questions

1. True or False: You will be interacting with the same specific group of other people throughout the experiment.
   The correct answer was "False."

2. True or False: You will be interacting with the same specific group of other people within any given period.
   The correct answer was "True."

3. The player in Role B chooses the payoff-reduction rule before he/she observes the allocation decisions of the people in Role A.
   The correct answer was "True."

4. If the player in Role B deducts tokens from a specific player in Role A, what happens to those tokens?
   The player in Role A gets to keep them
   The tokens are given to the other players in Role B
   The tokens simply disappear (correct)

5. Suppose you are in Role B and receive a message that the person in Role A with ID#1 "allocated his or her tokens to the common pot." Then you know that that person....
   Definitely allocated the tokens to the common pot
   May or may not have allocated the tokens to the common pot (correct)

6. Suppose you are in Role B and receive a message that the person in Role A with ID#1 "kept his or her tokens." Then you know that that person....
   Definitely kept his or her tokens
   May or may not have kept his or her tokens (correct)
7. Suppose all four people in Role A, including the person with ID #1, contributed 20 tokens to the common pot, for a total of 80 tokens. Suppose also that the person in Role B deducts 24 tokens from the payoffs of the person in Role A with ID #1, after choosing a "Small Investment." Then.....

...the payoffs of the person in Role A with ID #1 are:

20 + (0.4*80) = 52  
0 + (0.4*80) = 32  
0 + (0.4*80) - 24 = 8 (correct)  
0 + (0.4*20) - 24 = -16

8. ...and the payoffs of the person in Role B are:

20 + (0.4*80) = 52 (correct)  
20 + (0.4*80) + 24 = 76  
20 + (0.4*80) - 24 = 28  
20 + (0.4*80) - 2 = 50