Article

Journal of Conflict Resolution
2016, Vol. 60(7) 1312-1340

DOI: 10.1177/0022002714560349
jcr.sagepub.com

When Does Informal Enforcement Work?

Stine Aakre[1], Leif Helland[2], and Jon Hovi[1,3]

Abstract

We study experimentally how enforcement influences public goods provision when subjects face two free-rider options that roughly parallel the nonparticipation and noncompliance options available for countries in relation to multilateral environmental agreements (MEAs). Our results add to the MEA literature in two ways. First, they suggest that compliance enforcement will fail to enhance compliance in the absence of participation enforcement. Second, they indicate that compliance enforcement will boost compliance significantly in the presence of participation enforcement. Our results also add to the experimental literature on public goods provision, again in two ways. First, they reveal that previous experimental findings of enforcement boosting cooperation are valid only in settings with forced (or enforced) participation. Second, they show that subjects' willingness to allocate costly punishment points is significantly stronger when the enforcement system permits punishment of both types of free riding than when it permits punishment of only one type.

Keywords

international cooperation, international institutions, international regimes, game theory, international treaties

[1]Center for International Climate and Environmental Research, Oslo, Norway
[2]Department of Economics, BI Norwegian Business School, Oslo, Norway
[3]Department of Political Science, University of Oslo, Oslo, Norway

Corresponding Author:
Jon Hovi, Department of Political Science, University of Oslo, P.O. Box 1097 Blindern, Oslo 0317, Norway.
Email: jon.hovi@stv.uio.no

Can enforcement enhance cooperation on public goods provision? Particularly, can enforcement enhance cooperation in multilateral environmental agreements (MEAs) such as the Kyoto Protocol or the Montreal Protocol? If so, when is enforcement likely to work?

We seek answers to these questions by drawing on and contributing to two related yet very different literatures. The first is the MEA literature on the relevance and the significance of enforcement for compliance levels. In this literature, the enforcement school stresses the importance of coercive means[1] such as reciprocal measures,[2] financial penalties, trade restrictions, and suspension of privileges (Barrett and Stavins 2003; Downs, Rocke, and Barsoom 1996; Urpelainen 2011). In contrast, the managerial school maintains that "the effort to devise and incorporate [coercive measures] in treaties is largely a waste of time" (Chayes and Chayes 1995, 2) and instead advocates a facilitative approach based on capacity building, technical assistance, and transparency (Chayes and Chayes 1993, 1995).[3] These diverging views are rooted in contending ideas of what ultimately motivates states' behavior. While the enforcement school holds that behavior is guided by a "logic of consequentiality," the managerial school regards behavior as primarily motivated by a "logic of appropriateness."

Unsurprisingly, the two schools are widely regarded as mutually incompatible. In the words of Raustiala and Victor (1998, 681), they "reflect different visions of how the international system works, the possibilities for governance with international law, and the policy tools that are available and should be used to handle implementation problems." We argue that the two schools provide *complementary* insights concerning the relationships between enforcement, compliance, and participation. Whereas previous research has mostly considered compliance and participation *separately*,[4] we conducted an experiment enabling us to examine *jointly* the circumstances under which we should expect enforcement to induce countries to participate in and comply with MEAs. We find that the two schools are both right, but under different circumstances: compliance enforcement *fails* to enhance compliance without participation enforcement; however, it enhances compliance *substantially* with participation enforcement. We define participation enforcement as (positive or negative) incentives for countries to join (ratify) the agreement and refrain from withdrawing. Similarly, we define compliance enforcement as (positive or negative) incentives for participating countries to fulfill their commitments.

The second literature consists of laboratory experiments on public goods provision. This literature demonstrates that subjects' behavior in experiments deviates systematically from the predictions of standard game-theoretic models (i.e., models that assume a homogenous population of rational and purely self-regarding players). A main finding is that subjects are frequently prepared to punish uncooperative behavior in public goods experiments, even at a personal cost, and that such punishments are anticipated by subjects and therefore shape their behavior (e.g., Fehr and Gächter 2000; Kosfeld, Okada, and Riedl 2009; Ostrom 2000; Ostrom, Walker, and Gardner 1992). This finding would seem to challenge the managerial school's claim

that compliance enforcement is futile. However, whereas the managerial school is primarily concerned with the effect of compliance enforcement in MEAs without participation enforcement, practically all public goods experiments we know of implement *forced* participation (FP).

The experimental literature that uses a threshold public goods framework to study participation in MEAs constitutes an exception (e.g., Dannenberg 2012; Dannenberg, Lange, and Sturm 2014; McEvoy 2010). However, in line with much of the theoretical literature on participation in MEAs, this literature typically implements forced compliance, and hence ignores the fact that countries may join an MEA yet may choose to violate their commitments (to a larger or smaller degree). To our knowledge, Cherry and McEvoy (2013) and McEvoy et al. (2011) are the only previous experimental studies that allow both types of free riding (i.e., nonparticipation and noncompliance). However, their designs differ significantly from ours; in both studies, punishment for noncompliance is automatically carried out by a third-party enforcer.

In contrast, we report a three-stage experiment that implements voluntary participation (VP) and voluntary compliance. In our experiment, both participation enforcement and compliance enforcement are informal, in the sense that the players themselves decide not only the severity of the punishment but also what kind of behavior they want to punish. For example, participation does not entail a commitment to contribute a specific amount to the public good; thus, no objective contribution amount defines whether a participating subject is "compliant" or not.[5] We return to this point in our suggestions for future research in the conclusion.

In stage 1, subjects must choose whether to participate in a *project* that will benefit participants and nonparticipants equally. Stage 1, which is *not* included in most previous public goods experiments, represents countries' choice of whether to participate in an MEA (i.e., whether to ratify).

In stage 2, subjects choosing to participate in the project must decide how much of their fixed endowment they will contribute to the public good. Stage 2 represents MEA member countries' choice of a compliance level. While compliance is usually conceived of as the extent to which members of a treaty will meet their commitment (e.g., their target for emissions reductions), the subjects in our experiment make no explicit commitment. Thus, we use "compliance" in the somewhat looser sense of adherence to a group norm that develops endogenously. The idea is that subjects will likely accept contributions that exceed some minimum level and consider contributions below this minimum as unacceptable (and therefore as candidates for punishment).

In short, in our experiment, subjects can avoid making costly contributions in two ways—by not participating in the project or by participating without contributing. These two ways roughly parallel the two types of free riding available to countries in relation to MEAs: nonparticipation and noncompliance.

Finally, stage 3 is an enforcement stage, where subjects who have chosen to participate in the project (we henceforth refer to such subjects as "insiders") can

allocate punishment points to punish others. Punishment points are costly not only for the punished subject but also for the punishing insider. Our experiment implements four enforcement systems: no enforcement, only compliance enforcement, only participation enforcement, and both compliance enforcement *and* participation enforcement.

This design enables us to study the different enforcement systems' effect on the average participation in the project, on the average contribution *among insiders* (compliance), on the average *total* contribution (effectiveness), and on the extent to which insiders allocate punishment points.

Our results add to existing knowledge both in the MEA literature and in the experimental literature. Concerning the MEA literature, our results suggest that compliance enforcement *fails* to enhance compliance in the absence of participation enforcement; however, it enhances compliance *substantially* in the presence of participation enforcement. Thus, without participation enforcement, which is lacking in most existing MEAs,[6] designing institutions for compliance enforcement may well be futile, as argued by the managerial school. A major reason is self-selection; countries that are largely unwilling to contribute to problem solving will decline to participate in the first place. In contrast, for MEAs that include participation enforcement (e.g., by permitting member countries to punish nonmembers), designing effective institutions for compliance enforcement may be essential, as argued by the enforcement school. Constructing effective MEAs for particularly challenging problems, such as climate change, may well require use of incentives that induce reluctant countries to participate. Hence, our experimental results support both schools' arguments but under different circumstances.

More generally, our results suggest that enforcing only compliance or only participation will likely have little bearing on the effectiveness (measured by the average total contribution) of MEAs that aim to provide a public good. Compliance enforcement without participation enforcement will cause free riding to take the form of nonparticipation, whereas participation enforcement without compliance enforcement will cause free riding to take the form of noncompliance. To deter both types of free riding, MEAs must enforce *both* participation *and* compliance.

With respect to the experimental literature, we find that both contribution levels and allocation of costly punishment points depend on the institutional setting. Permitting insiders to punish either only other insiders (compliance enforcement) or only outsiders (participation enforcement) has little or no effect on the average total contribution; in particular, enabling insiders to punish other insiders has no noticeable effect when outsiders cannot be punished. In contrast, enabling insiders to punish *both* other insiders *and* outsiders enhances the average total contribution substantially.

The institutional setting's effect on the average total contribution is paralleled by its effect on the allocation of punishment points. *More* free riding takes place when an escape option exists (i.e., when the MEA enforces only participation or enforces only compliance) than when no such option exists. Nevertheless, insiders allocate

fewer punishment points when an escape option exists, presumably because the escape option makes punishment futile. Thus, although the motivation for allocating punishment points clearly cannot be fully rational *and* purely self-regarding, our results suggest that it does have an instrumentally rational component: insiders allocate costly punishment points primarily when the punished subject cannot avoid further punishment by switching to a different form of free riding.

International relations scholars have thus far made only limited use of experimental methods. Given our desire to explore different enforcement systems' effects on countries' behavior, an experimental approach seems particularly well suited (see, e.g., McDermott 2011; Tingley 2011). Although real-world examples of each type of enforcement system do exist, most MEAs lack enforcement provisions. Nevertheless, MEAs often exhibit near-perfect compliance levels. Such lack of variation hardly permits use of field data to properly test whether enforcement enhances compliance, participation, or both. In addition, an experimental design permits careful control and manipulation of the variables of central interest.

We study a highly stylized environment: the game's structure is public knowledge; time periods, the time horizon, punishments, contributions, and participation are all unambiguously defined; subjects can observe behavior instantaneously and without noise; and interaction is anonymous. Clearly, this stylized environment bears only slight resemblance to real-world MEA settings. Thus, the external validity of our experimental findings should be checked through further studies based on field data—if and when relevant field data become available.

This being said, choosing an experimental study based on a stylized environment has its advantages. In particular, it permits us to carefully tailor a design for studying whether and when controlled and truly exogenous institutional variation influences participation, compliance, and punishment behavior.

The next section provides a review of both the MEA literature debating whether enforcement enhances compliance and the experimental literature on public goods provision. Then, having provided one MEA example for each of the four enforcement systems considered in our experiment, we outline our model, discuss our experimental design, and present our results. In the final section, we present our conclusions.

Previous Research

We first review the MEA literature debating the relevance and significance of enforcement for compliance levels, and then the experimental literature on public goods provision.

The MEA Literature on the Relevance and Significance of Enforcement

Both the managerial school and the enforcement school start from the premise that—because of the anarchical nature of the international system—countries cannot

guarantee to honor their commitments (Axelrod and Keohane 1986; Oye 1985). Hence, both schools see it as essential to map features that permit countries to bind themselves to mutually beneficial courses of action, and both schools find it essential to identify strategies that might enhance cooperation.[7] They also agree that MEA compliance has generally been high[8] and that enforcement has apparently played little or no role in achieving that record (Brown Weiss and Jacobson 1998; Chayes and Chayes 1993; Downs, Rocke, and Barsoom 1996).

However, the two schools disagree on at least three matters (Tallberg 2002). First, they disagree on whether enforcement influences compliance. The managerial school considers enforcement to be largely irrelevant and argues that states' "general propensity to comply" with MEAs is due more to efficiency, national interests, and regime norms than to states' concerns with enforcement (Chayes and Chayes 1995). In contrast, the enforcement school contends that compliance in deep MEAs requires enforcement measures that offset the benefits a state could obtain by not complying.[9] It argues that widespread compliance despite little enforcement is only to be expected, given that states are generally reluctant to accept obligations they are unable or unwilling to meet.[10] Thus, MEAs are often shallow, in the sense that they commit countries to little more than they would be prepared to do unilaterally, and such shallow MEAs entail little incentive for noncompliance.[11] According to the enforcement school, it would be a mistake to infer from high compliance to *shallow* MEAs without enforcement that *deep* MEAs can also achieve high compliance without enforcement.

Second, the two schools differ in their interpretation of those relatively few instances of noncompliance that *are* observed. For the managerial school, cases of noncompliance are usually *not* attempted free riding through deliberate defiance of the legal standard. Rather, such cases are caused by (1) the ambiguity and indeterminacy of treaties, (2) the limited capacity of states to comply, and (3) social and economic changes due to time lags between commitments and their implementation. In contrast, the enforcement school argues that the causes of noncompliance are to be found in the incentive structure: states choose to be noncompliant when the benefits of noncompliance exceed the costs of being detected and punished.

Finally, these two sources of disagreement have implications with respect to what the two schools see as potential remedies for avoiding noncompliance and for reestablishing compliance. According to the enforcement school, "a punishment strategy is sufficient to enforce a treaty when each side knows that if it cheats it will suffer enough from the punishment that the net benefit will not be positive" (Downs, Rocke, and Barsoom 1996, 385). In contrast, the managerial school argues that noncompliance is better addressed by (1) improving dispute resolution procedures, (2) supplying technical and financial assistance, and (3) increasing transparency. Thus, the managerial school regards regimes as playing an "active role . . . in modifying preferences, generating new options, persuading the parties to move toward increasing compliance with regime norms, and guiding the evolution of the normative structure in the direction of the overall objectives of the regime" (Chayes and Chayes 1995, 229).

To summarize, the managerial and enforcement schools disagree sharply on whether compliance enforcement matters in MEAs. Our experiment suggests that the answer to this question depends on whether participation is also enforced.

The Experimental Literature on Public Goods Provision

The experimental literature on public goods provision largely considers variations of the following game:[12] n subjects endowed with z units of a *numéraire* good decide simultaneously how much of their endowment they will keep for themselves and how much they will contribute to a public account for the subject group. Contributions are multiplied by a factor weakly greater than one and strictly less than n and are divided equally among all n subjects. This game's unique Nash equilibrium is that all subjects keep their entire endowment for themselves. However, this equilibrium is Pareto suboptimal; if all subjects were to contribute their entire endowment to provision of the public account, every subject would be better off.

Some experiments add an enforcement stage,[13] wherein subjects can punish other subjects by allocating punishment points. Typically, one allocated punishment point detracts three units of the *numéraire* good from the punished player's payoff. However, it also detracts one unit of the *numéraire* good from the punishing player's payoff: so punishing is costly. If all players are rational and purely self-regarding, the subgame-perfect equilibrium of this extended game (with enforcement) is thus that all subjects keep their entire endowment and that no punishment takes place.

Many experimental studies of such public goods games exist (e.g., Fehr and Gächter 2000, 2002; Kosfeld, Okada, and Riedl 2009). Typically, these experiments allow subjects to play the game a fixed number of times (often ten). The subgame-perfect equilibrium in such a repeated game (still assuming rational and purely self-regarding players) is that no player ever contributes any of its endowment and that no player is ever punished.

However, in experiments, the behavior predicted by the subgame-perfect equilibrium of the standard model is generally not observed. If allocation of punishment points is *not* possible, average contributions typically start at sizable levels (40–50 percent of the endowment) in the first period and gradually taper off, reaching a fairly low level (10–15 percent of the endowment) by the last period. In contrast, if allocation of punishment points *is* possible, average contributions typically start at a higher level (60–70 percent of the endowment), and approach maximum levels (90–100 percent of the endowment) by the last period. Thus, the possibility of allocating punishment points influences behavior in forceful ways. While the exact mechanisms producing these results are not well understood, it is generally thought that heterogeneity in subject motivations plays an important role. A sizable portion of subjects seem to be "reciprocators" (see, e.g., Fehr, Fischbacher, and Gächter 2002). Reciprocators increase (decrease) their current contribution if their contribution in the preceding period was below (above) the average contribution in the rest of their group. The observed decline in average contributions in the no-punishment

treatment is thought to stem from reciprocators' being overoptimistic concerning the subject mix; in particular, reciprocators tend to underestimate the portion of purely self-regarding players. Numerous replications suggest that purely self-regarding subjects constitute roughly one-third of the subject pool (in modern societies). Reciprocators' overoptimistic bias concerning the subject mix causes them to provide sizable contributions in the first period and to adjust downward over time, as they observe average contributions below their expectations.[14]

When allocation of punishment points is possible, subjects can discipline free riders. Although allocating punishment points is costly, subjects often allocate them. It is generally thought that "strong reciprocity" plays a role here. A strong reciprocator is prepared to sacrifice material gains to punish subjects that violate cooperative social norms.[15] If a purely self-regarding player believes strong reciprocation is sufficiently widespread, and if allocation of punishment points is possible, contributing at a level that avoids punishment may well be a best response (see, e.g., Fehr and Fischbacher 2005; Gürerk, Irlenbusch, and Rochenbach 2006).

A large literature on endogenous group selection exists (e.g., Ahn, Isaac, and Salmon 2009; Charness and Yang 2010). In this literature, subjects self-select into groups, and only group members benefit from goods production; hence, this literature addresses provision of a club good. In contrast, we focus on provision of a pure public good (such as mitigation of climate change), meaning that both members (insiders) and nonmembers (outsiders) benefit from public goods production.

As mentioned in the introduction, very few previous public goods experiments implement VP. The relative lack of such experiments is remarkable, considering that in many real-world social situations (including international treaty making) decision makers clearly have a choice between participating and not participating. Our experiment shows that implementing VP may influence results substantially, depending on the institutional setting.

Examples

Relatively few MEAs include enforcement provisions; however, some do. To give the reader an idea of what the various enforcement systems might look like in practice, we now offer one MEA example for each of the four systems in our experiment.

First, the 1985 Helsinki Protocol lacks provisions for enforcement of either compliance or participation. Helsinki requires signatories to "reduce their national annual sulphur emissions or their transboundary fluxes by at least 30 per cent as soon as possible and at the latest by 1993, using 1980 levels as the basis for calculation of reductions" (Article 6). Despite the lack of enforcement provisions, all signatories achieved their emissions reduction targets by the set deadline. However, participation was limited; for example, major emitters such as the United States and the United Kingdom declined to ratify the agreement. Moreover, incentives for noncompliance were also limited because Helsinki merely codified what the parties were

prepared to do unilaterally (Bratberg, Tjøtta, and Øines 2005; Ringquist and Kostadinova 2005).

Second, the enforcement system of the Kyoto Protocol aims to enforce compliance but not participation. Participation in Kyoto is far from full; of the 192 countries that ratified Kyoto, only 36 had emissions limitation targets during Kyoto's first commitment period, and for countries such as Russia and Ukraine, the targets were not effective (the so-called hot air problem).[16] Furthermore, Kyoto's compliance enforcement system has been severely criticized. One of many objections is that a noncompliant member country can avoid punishment by withdrawing (Barrett 2003); such avoidance is possible because Kyoto does not enforce participation. Kyoto's first commitment period expired on December 31, 2012. Whether some countries participated without complying (fully) with their targets is not known at the time of writing.

Third, the Montreal Protocol's *first* enforcement system attempted to enforce participation but not compliance. That system allowed member countries to impose restrictions on trade with nonmembers in substances that threaten the ozone layer. Anecdotal evidence suggests that this enforcement system induced some countries to participate. There is "direct evidence from some countries that the trade provisions were important in persuading them to accede to the treaty; a good example is the Republic of Korea, which initially expanded its domestic CFC production, but realizing the disadvantages of being shut out of Western markets, became a party" (Brack 2003, 220).

Finally, the Montreal Protocol's current (second) enforcement system seeks to enforce both compliance and participation.[17] Montreal is widely regarded as highly successful. Several factors contribute (singly or in combination) to this success; one might be Montreal's enforcement system. While the first compliance system enforced only participation, provisions for enforcing compliance were soon added. Again, anecdotal evidence suggests that Montreal's compliance enforcement system has induced some members to fulfill their commitments. In several cases, the use of measures from the indicative list of consequences has been threatened against noncompliant members: "Their use has been threatened, in a series of MOP decisions, usually in the following terms: 'These measures may include the possibility of actions available under Article 4, such as ensuring that the supply of CFCs . . . is ceased and that exporting parties [parties exporting to the non-complying party] are not contributing to a continuing situation of non-compliance.' So far, this provision has never had to be used, but, as with the former non-parties that decided to accede, its existence appears to be important in encouraging compliance" (Brack 2003, 220).

Model

Consider a three-stage one-shot game with n players, each of whom is endowed with z units of a *numéraire* good. In stage 1, all players decide simultaneously whether to

participate in a project. Participating reduces a player's endowment from z to $z(1 - d)$ units, where $0 < d < 1$; hence, participation is costly.[18] Once made, the participation decisions become public knowledge for the n players. We denote players who participate "insiders" and players who do not participate "outsiders." Let m be the number of insiders, so that $n - m$ is the number of outsiders.

In stage 2, insiders decide simultaneously how much of their endowment they will contribute to a public good; thus, insider i's contribution must satisfy $c_i \in [0, z(1 - d)]$. Outsiders cannot contribute; hence, for outsiders $c_i = 0$. Once made, the contribution decisions become public knowledge for the n players.[19]

Stage 3 is an enforcement stage. We consider the effect of three enforcement systems for a regime with VP:

VP Treatment 1: Insiders can punish other insiders.
VP Treatment 2: Insiders can punish outsiders.
VP Treatment 3: Insiders can punish both insiders and outsiders.

As a control, we also consider the effect of enforcement for a regime with FP.

In all three VP treatments, as well as in the FP treatment, subjects play ten periods *without* enforcement, followed by ten periods *with* enforcement.[20] Hence, our design enables us not only to compare behavior across different enforcement systems but also to compare behavior under each enforcement system to behavior *without* enforcement.

Regardless of the enforcement system, outsiders cannot punish other players. Punishment consists of insider i allocating punishment points ($PP \in [0, \overline{PP}]$) to a punishable player j. PP_{ij} denotes the number of punishment points allocated by i to j, while PP_{ji} denotes the number of punishment points allocated by j to i. We consider the usual punishment technology in which one PP detracts one unit of the *numéraire* good from the *punishing* player's payoff and detracts three units of the *numéraire* good from the *punished* player's payoff. Once made, punishment decisions become public knowledge for the n players, payoffs are distributed, and the game ends.

If player i is an insider, its payoff U_i^I is as follows:

$$U_i^I = z(1 - d) - c_i + \frac{\alpha}{n}\sum_{j=1}^{m} c_i - \sum_{j=1}^{n} PP_{ij} - 3\sum_{j=1}^{m} PP_{ji}.$$

Here the first term represents insider i's endowment, the second term represents its contribution, the third term represents its benefit from own and other insiders' contributions (with α/n representing the marginal return of a unit contributed to the public good), the fourth term represents the cost i incurs by punishing other players, and the fifth term represents the cost it incurs by being punished by other insiders. The fifth term equals zero if the enforcement system does not permit punishment of insiders and if the system permits such punishment but no other insider chooses to punish insider i.

Similarly, if player i is an outsider, its payoff U_i^0 is as follows:

$$U_i^0 = z + \frac{\alpha}{n}\sum_{j=1}^{m} c_j - 3\sum_{j=1}^{m} PP_{ji}.$$

Here the first term represents outsider i's endowment, the second term represents its benefit from insiders' contributions, and the third term represents the cost it incurs by being punished by insiders. This third term equals zero if the enforcement system does not permit punishment of outsiders and (when punishment of outsiders *is* permitted) if no insider chooses to punish outsider i. Our design satisfies $\frac{1}{1-d} < \alpha < n$.

Consider a situation in which it is common knowledge that all n players are rational and purely self-regarding. What will be the game's subgame-perfect equilibrium? Using backward induction, we find that if stage 3 is reached, no insider will punish, because such punishment is costly for the punishing player. If stage 2 is reached, no insider will make a contribution, because the insider's marginal cost of contributing one unit of the *numéraire* good is 1; in contrast, its marginal benefit of contributing one unit is $\frac{\alpha}{n} < 1$. Finally, at stage 1, no player will participate, because insiders' endowment is $z(1 - d)$, whereas outsiders' endowment is $z > z(1 - d)$. Hence, for all treatments, the unique subgame-perfect equilibrium is that all n players choose to be outsiders, meaning that in equilibrium, stages 2 and 3 are never reached and each player's payoff equals its endowment z. Provided that $\frac{1}{1-d} < \alpha$ (which holds in our design), this subgame-perfect equilibrium is Pareto dominated by the nonequilibrium outcome wherein (1) all players participate, (2) all players contribute their entire endowment $z(1 - d)$, and (3) no player allocates punishment points.[21]

Backward induction shows that in a finitely repeated game such as ours, the stage-game equilibrium will be played in every period.

Design and Implementation

Our experiment used the following parameter values: $n = 4$, $z = 22$, $d = \frac{1}{11}$, and $\alpha = 1.6$. Each insider could allocate integers from zero to ten as punishment points to each punishable player in its group. We implemented the following rules: (1) if a subject's net income after stage 3 was positive, the subject received that (positive) income for that period; (2) if a subject's net income after stage 3 was negative, we limited the subject's loss for that period to the cost of the punishment points the subject *allocated* in that period. To make bankruptcies unlikely, we allocated twenty-five additional units of the *numéraire* good to each subject after period 10.[22]

As explained in the previous section, we ran three VP treatments and one FP treatment (as a control). The first ten periods of each VP treatment consisted of stages one and two only; that is, they included no enforcement. The last ten periods

consisted of stages 1, 2, and 3; that is, these periods included enforcement. However, the enforcement system differed across the VP treatments.

In the FP treatment, $z = 20$ and $d = 0$. In this treatment's first ten periods, only stage 2 was played, so there was no enforcement. In the last ten periods, stages 2 and 3 were played, which enabled every group member to punish other group members.[23] In all other respects, the FP treatment was identical to the three VP treatments.

We recruited 180 subjects from among bachelor students at the second author's institution. In each of the three VP treatments, forty-four subjects participated (in groups of four); hence, each VP treatment included eleven groups. In the FP treatment, forty-eight subjects participated (again in groups of four); hence, the FP treatment included twelve groups.

We implemented a partner matching in which the four-subject groups were formed randomly at the beginning of each treatment and remained constant for that treatment's twenty periods. In each period, subjects received feedback on contributions and punishments made by other subjects in their own group. This feedback was provided in separate columns. However, to prevent the possibility of reputation building, subjects were randomly rotated over columns from one period to the next. These facets of our design permitted learning and adjustment in constant groups, while making each period very similar to a one-shot game. Our design also facilitated implicit in-group coordination.[24] Subjects' anonymity was preserved throughout.

We programmed the experiment in *z*-tree, using Herrmann, Thöni, and Gächter's (2008) *z*-tree programming, computer screens, and instructions as a point of departure.[25] We modified their programming, computer screens, and instructions only when required to accommodate our various treatments.

In each treatment, the administrator, having seated the subjects at randomly drawn cubicles in the lab, distributed the instructions and read them aloud. The treatment began after subjects had answered a set of control questions designed to ensure they understood the payoff structure. Each treatment lasted about two hours.

Subjects received a show-up fee of US$20 in addition to whatever they earned in the experiment. They received their earnings, which averaged around US$40, in cash and privately, at the end of the treatment concerned.[26]

Results

Here we compare the results from our four treatments with respect to participation levels, average contributions *among insiders*, average *total* contributions, and allocated punishment points. Compared to traditional experimental designs that implement FP, our design with VP entails novel and interesting findings that throw new light on the debate between the enforcement school and the managerial school.

Given our matching protocol, subject-level observations are independent across groups, but not within groups (where strategic interaction took place). We therefore use within-treatment group averages over all periods as data points.[27] We test between-group effects (differences across treatments within the nonpunishment and

punishment phases, respectively) using the Wilcoxon two-sample rank-sum test. We test within-group effects (differences between nonpunishment and punishment phases for a given treatment) using the Wilcoxon matched-pairs rank-sum test. We also report a set of regressions to check our results' robustness.

Participation

When only insiders are punishable, would-be free riders can escape punishment by being outsiders.[28] This escape option influences public goods provision negatively by limiting the number of insiders:

Result 1: The average number of insiders is lower in VP treatment 1 (only insiders punishable) than in VP treatments 2 (only outsiders punishable) and 3 (both insiders and outsiders punishable).

In the FP treatment, subjects *cannot* choose to be outsiders. Hence, in this treatment, participation remains constant at 100 percent for all twenty periods. In contrast, in the three VP treatments, subjects *can* choose to be outsiders. Here, average participation varies considerably across treatments and across periods but never comes close to 100 percent. In all three VP treatments, average participation starts at around 70 percent in period 1 and gradually tapers off to around 45 percent in period 10 (Figure 1). From period 11, VP treatments 2 and 3 introduce participation enforcement (insiders can punish outsiders). In these two treatments, average participation then rises sharply to between 80 and 90 percent and largely remains there for the last ten periods. In contrast, VP treatment 1 does not introduce participation enforcement (from period 11, insiders can punish other insiders but not outsiders). Here average participation continues to taper off during the last ten periods, ending at 20 to 25 percent.

We find that in periods 11 through 20, the average participation level is significantly lower in VP treatment 1 than in VP treatment 2 ($p_{\mathrm{VP1,VP2}} = .001$) and VP treatment 3 ($p_{\mathrm{VP1,VP3}} = .001$).

Average Contribution among Insiders

When only outsiders are punishable, would-be free riders can escape punishment by participating without contributing. Although participation is costly, so that would-be free riders must weigh the participation cost against the cost of being punished for nonparticipation, this escape option causes high participation rates while undermining public goods provision through low average contributions among insiders:

Result 2: The average contribution among insiders is lower in VP treatments 1 (only insiders punishable) and 2 (only outsiders punishable) than in VP treatment 3 (both insiders and outsiders punishable).

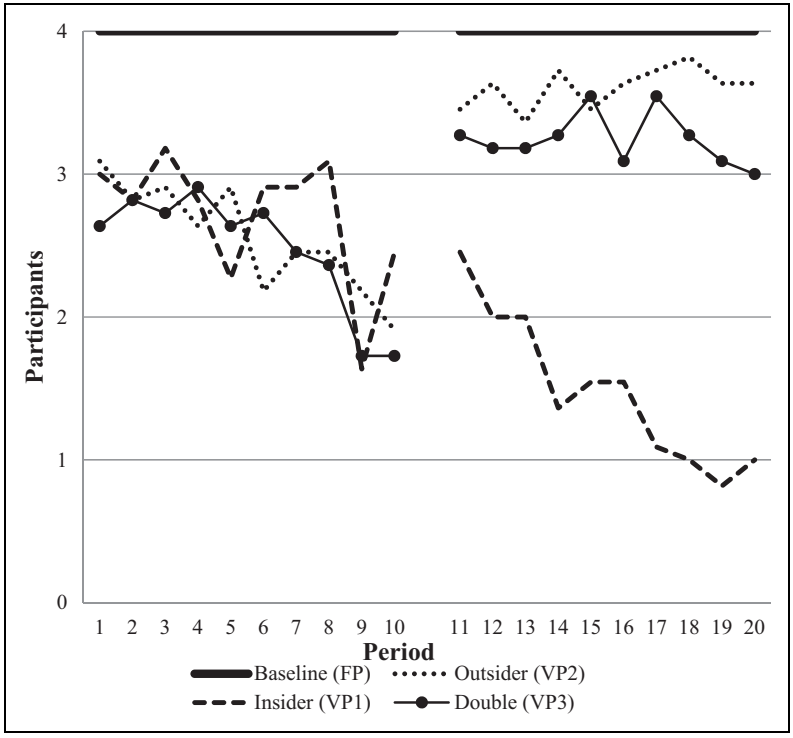

Figure 1. Average participation by treatment and period.

In the FP treatment, the average contribution among insiders starts 55 between 60 percent and then gradually tapers off to between 20 and 25 percent in period 10.[29] In the three VP treatments, the average contribution starts somewhat lower (between 35 percent and 50 percent), but here too it gradually tapers off, reaching 10 to 25 percent by period 10 (Figure 2). In periods 1 through 10, no significant difference exists between the three VP treatments ($p_{VP1,VP2} = .870$; $p_{VP1,VP3} = .375$; $p_{VP2,VP3} = .670$), and only one of them (VP treatment 3) differs significantly from the FP treatment ($p_{FP,VP1} = .140$; $p_{FP,VP2} = .132$; $p_{FP,VP3} = .010$).

From period 11, the average contribution among insiders in the FP treatment rises sharply (to 60 percent) and stays at 60–75 percent for the remaining periods. VP treatment 3 displays a similar effect. By contrast, in VP treatments 1 and 2, the average contribution among insiders remains at roughly the same level in periods 11 through 20 as in periods 1 through 10. Thus, the average contribution among insiders is significantly lower in VP treatments 1 and 2 than in VP treatment 3 ($p_{VP1,VP3} = .011$; $p_{VP2,VP3} = .005$) in the last ten periods. No statistically significant difference exists between VP treatments 1 and 2 ($p_{VP1,VP2} = .718$).

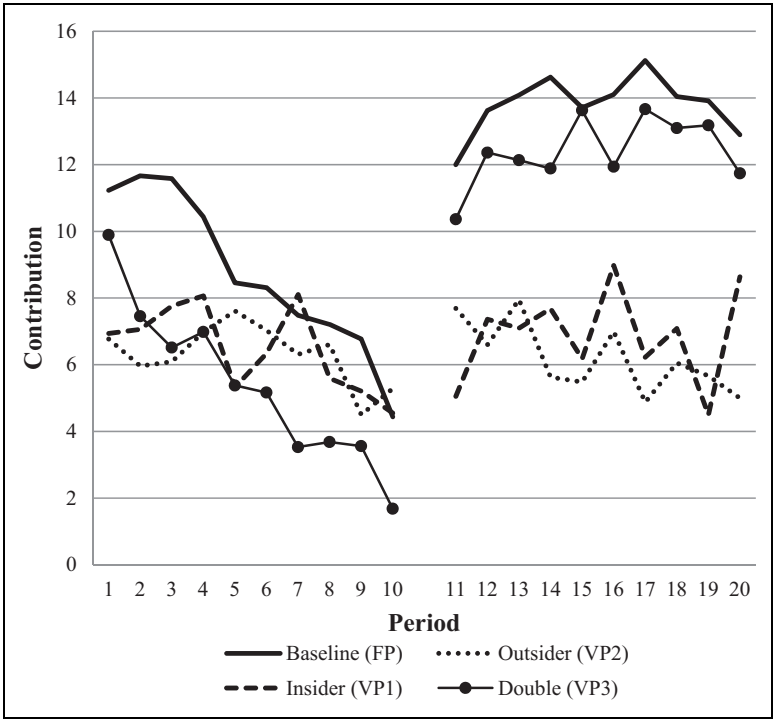

Figure 2. Average contribution among insiders by treatment and period.

Why are insiders' contributions low in VP treatment 1? At low participation rates (few insiders), exploitation of insiders by free-riding outsiders is substantial. However, insiders cannot correct such free riding through directed punishment; indeed, the only option insiders have for hurting free-riding outsiders is to limit their contributions.

Concerning the debate between the enforcement school and the managerial school, particularly interesting to note is that making only insiders punishable has no discernible effect on the average contribution among insiders; for VP treatment 1, no statistically significant difference exists between the initial ten periods without enforcement and the last ten periods with enforcement ($p_{\mathrm{VP1}} = .657$). This finding supports, for cases without participation enforcement, the managerial school's claim that compliance enforcement is largely pointless in MEAs.

Average Total Contribution

Recall that an escape option exists in VP treatments 1 and 2 but not in VP treatment 3 or in the FP treatment. Our results suggest that when would-be free riders have an escape option, enforcement does *not* enhance the average total contribution:

Result 3: In VP treatments 1 and 2, the average total contribution in the last ten periods is not significantly higher than in the first ten periods.

In contrast, when would-be free riders do *not* have an escape option, enforcement enhances the average total contribution substantially:

Result 4: In VP treatment 3 and in the FP treatment, the average total contribution in the last ten periods is significantly higher than in the first ten periods.

In the FP treatment, participation invariably equals 100 percent; hence, the average total contribution necessarily equals the average contribution among insiders. As we have seen, this level starts around 55 percent, gradually tapers off to around 25 percent by period 10, increases sharply to between 65 and 70 percent when allocation of punishment points becomes possible from period 11, and stays roughly at that level for the last ten periods.

Unsurprisingly, as the participation level in the three VP treatments is invariably below 100 percent, the average total contribution in those treatments begins much lower (around 33 percent). Here too, the average total contribution gradually tapers off, reaching 10 to 15 percent by period 10 (Figure 3).

From period 11, the observed pattern differs sharply between the three VP treatments. In VP treatment 1 (only insiders punishable), the average total contribution continues to taper off to around 10 percent by period 20. In VP treatment 2 (only outsiders punishable), the average total contribution increases slightly from period 10 to period 11, but then starts tapering off again.

Perhaps the most important finding from VP treatments 1 and 2 is that enforcing *either* only compliance (VP treatment 1) *or* only participation (VP treatment 2) does *not* increase the average total contribution. We find that in VP treatment 2, the average total contribution in the last ten periods is not significantly different from the level in the first ten periods ($p_{VP2} = .182$). Concerning VP treatment 1, we find that the average total contribution in the last ten periods is actually significantly *lower* than in the first ten periods ($p_{VP1} = .041$).[30]

However, enforcing *both* participation *and* compliance has a very substantial positive effect on the average total contribution. In VP treatment 3 (both insiders and outsiders punishable), the average total contribution increases sharply from period 10 to period 11 and remains high for the last ten periods. Both in VP treatment 3 and in the FP treatment, the average total contribution in the last ten periods is significantly higher than in the first ten periods ($p_{VP3} = .004$, $p_{FP} = .008$).

Testing for differences in average total contributions across treatments reveals the following pattern: in periods 1 through 10, VP treatments 1 through 3 are not significantly different from each other; however, they are all significantly different from the FP treatment.[31] In periods 11 through 20, the FP treatment, and the VP3 treatment are not significantly different from each other; however, each of these two treatments is significantly different both from the VP1 treatment and from the

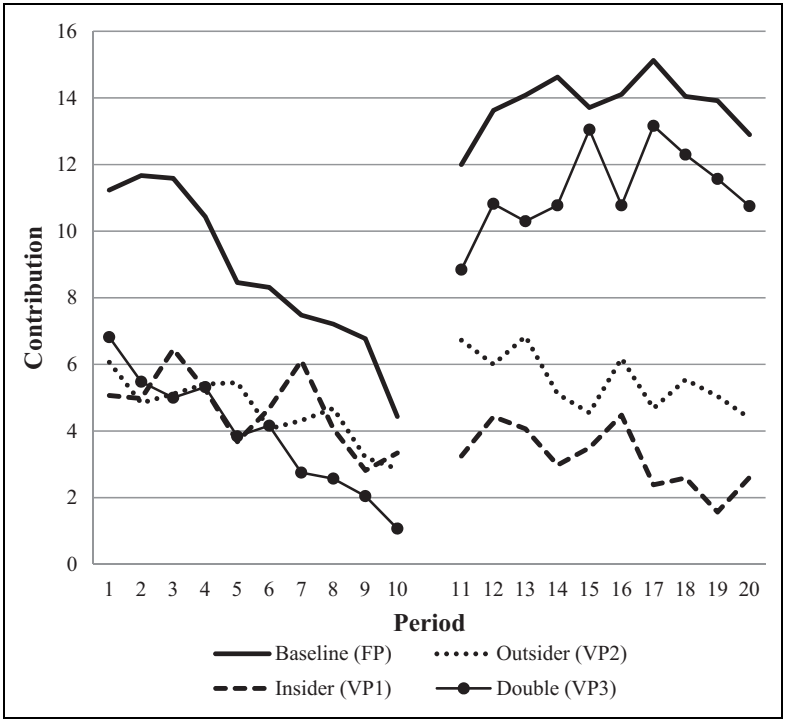

Figure 3. Total average contribution by treatment and period.

VP2 treatment. Finally, a significant difference between the VP1 treatment and the VP2 treatment exists in the final ten periods.[32] The latter result suggests that enforcing only compliance is somewhat more effective than enforcing only participation (although both of these two enforcement regimes are rather ineffective).

To explore the robustness of our findings from the nonparametric tests concerning total group contributions (see Figure 3), we now present a set of regressions that exploit our data's panel structure. First, we ran separate regressions for periods 1 through 10 and for periods 11 through 20. For each set of periods, total group contributions were regressed on treatment dummies and period dummies, using robust standard errors (*SE*s) clustered on the forty-five unique groups in the experiment. Tables 1 and 2 show the results.

For periods 1 through 10 (Table 1), the three VP treatments' group contribution lies 16 to 19 experimental currency units (ECUs) below that of the FP treatment. The coefficients for all of the three treatment dummies are significantly different from zero at conventional levels; in contrast, *t*-tests suggest that the null hypothesis of no difference between the regression coefficients cannot be rejected for any pair of VP treatments.[33] Furthermore, the period dummies show a negative trend in group

Table 1. Treatment Effects (reference category = FP) with Period Dummies (reference period = 1).

			95 percent CI	
	Coefficient	*SE*	Low	High
Constant	42.4***	4.64	33.0	51.7
VP1 (insider)	−16.5***	4.88	−26.3	−6.6
VP2 (outsider)	−16.6***	5.34	−27.4	−5.8
VP3 (both)	−19.4***	4.74	−28.9	−9.8
*t*2	−2.1	2.16	−6.5	2.2
*t*3	−1.0	2.81	−6.6	4.7
*t*4	−2.8	2.26	−7.3	1.8
*t*5	−7.8***	2.38	−12.6	−3.0
*t*6	−8.0***	3.05	−14.2	−1.9
*t*7	−8.6***	3.00	−14.7	−2.6
*t*8	−10.8***	3.10	−17.0	−4.5
*t*9	−14.4***	2.85	−20.2	−8.7
*t*10	−17.7***	3.17	−24.1	−11.3
$F(12,44)$	8.5***			
R^2	.28			
N	450			

Note: CI = confidential interval; *SE* = standard error. Robust *SE*s clustered on forty-five unique groups. Dependent: total group contributions.
Significance level: ***1 percent. **5 percent. *10 percent.

contributions over time; however, this trend is significantly different from zero only from period 5 onward.

For periods 11 through 20, too (Table 2), the three VP treatments' group contribution lies below that of the FP treatment. The coefficients for VP treatments 1 and 2 are significantly different from zero; in contrast, the coefficient for VP treatment 3 is *not* significantly different from zero at conventional levels. Furthermore, a *t*-test permits us to reject the null hypothesis of no difference between the regression coefficients of VP treatments 1 and 3. The same holds for the coefficients of VP treatments 2 and 3 but *not* for the coefficients of VP treatments 1 and 2.[34] The period dummies show a *positive* trend in group contributions over time; however, this trend is not significantly different from zero (after period 12).

Second, we ran a regression to check—treatment by treatment—whether total group contributions differ between periods 11 through 20 and periods 1 through 10 (Table 3). Total group contributions were regressed on a set dummy (scores one if the period number is greater than ten, zero otherwise) and a time trend. Again, robust *SE*s are clustered on groups.

The set dummy coefficient is significantly different from zero both for the FP treatment and for VP treatment 3. It is also significantly different from zero for VP treatment 2; however, the set dummy coefficients for the FP and VP3 treatments

Table 2. Treatment Effects (reference category = FP) and Period Dummy Effects (reference period = 11).

			95% CI	
	Coefficient	*SE*	Low	High
Constant	52.2***	4.49	43.2	61.3
VP1 (insider)	−42.5***	4.98	−52.5	−32.4
VP2 (outsider)	−33.2***	5.49	−44.3	−22.1
VP3 (both)	−10.3	7.92	−26.3	5.6
*t*12	4.1*	2.34	−.6	8.8
*t*13	4.6	2.78	−1.0	10.2
*t*14	2.8	2.64	−2.5	8.2
*t*15	4.0	2.87	−1.7	9.8
*t*16	4.8	3.14	−1.5	11.1
*t*17	4.7	3.15	−1.6	11.1
*t*18	3.7	3.38	−3.1	10.6
*t*19	1.4	3.72	−6.1	8.9
*t*20	−.8	3.81	−7.7	7.6
$F(12,44)$	12.1***			
R^2	.46			
N	450			

Note: CI = confidential interval; *SE* = standard error. Robust *SE*s clustered on forty-five unique groups. Dependent: total group contributions.
Significance level: ***1 percent. **5 percent. *10 percent.

Table 3. Treatment Effects for Periods 1–10 versus 11–20.

	FP (baseline)	VP1 (insider)	VP2 (outsider)	VP3 (both)
Constant	42.7*** (5.83)	23.4*** (2.48)	23.8*** (4.15)	19.5*** (4.17)
Set dummy	34.1*** (6.52)	3.0 (3.04)	13.4*** (4.13)	36.3*** (7.37)
Time	−1.4** (.51)	−.9** (.31)	−1.0* (.51)	−.7 (.62)
No. of groups	12	11	11	11
No. of subjects	48	44	44	44
F	13.9*** $F(2,11)$	5.54** $F(2,10)$	7.11** $F(2,10)$	14.27*** $F(2,10)$
R^2	.25	.08	.04	.34

Note: Regression coefficients (robust standard errors clustered on unique groups). Dependent: total group contributions.
Significance level: ***1 percent. **5 percent. *10 percent.

are much larger than the set dummy coefficient for VP treatment 2. Interestingly, the absolute effect of introducing enforcement is even stronger in VP treatment 3 than in the FP treatment, even though the total group contributions start from a lower level in VP treatment 3. The time trend is negative for all treatments and is (marginally)

statistically significant for the first three. R^2 indicates that the model specification shown in Table 3 explains the dependent variable's variance in the FP and VP3 treatments substantially better than it explains the corresponding variance in VP1 and VP2.

Finally, we ran regressions to check—treatment by treatment—the evolution of total group contributions across periods in the nonpunishment and punishment phases, respectively (see Table S1 in the supplementary material). In all treatments, total per-period contributions drop substantially when punishments are unavailable (periods 1–10). In all treatments, the null hypothesis that contributions are not decreasing is rejected at conventional levels. When punishments are available (periods 11–20), total per-period contributions increase in the FP and VP3 treatments, but not significantly. In contrast, total per-period contributions continue to fall in VP treatments 1 and 2, where we reject the null hypothesis that contributions are not decreasing, using a one-sided test. Testing for differences between the nonpunishment and punishment period coefficients for each treatment yielded the following results (see Table S2 in the supplementary material): no significant difference exists in VP treatments 1 and 2; in contrast, the difference between the coefficient for periods 1 through 10 and the coefficient for periods 11 through 20 is highly significant in the FP and VP3 treatments (z-test, $p < .01$).

In sum, our regressions support the results of the nonparametric analysis reported earlier. In particular, they demonstrate the robustness of results 3 and 4, which essentially say that enforcing both participation and compliance is necessary to enhance contribution levels substantially.[35]

Allocation of Punishment Points

As we have seen, the average total contribution is modest in VP treatments 1 and 2; plenty of free-riding activity takes place. Nevertheless, subjects allocate very few punishment points. In both treatments, the average number of allocated punishment points starts as low as 1 to 1.5 (period 11) and ends even lower (period 20). This pattern seems to reflect the futility of punishing outsider free riding when insider free riding cannot be punished and the futility of punishing insider free riding when outsider free riding cannot be punished.

In contrast, the average total contribution in the FP treatment and in VP treatment 3 is significantly higher; considerably *less* free riding takes place. Nevertheless, the average number of allocated punishment points is *greater*—more than twice the number allocated in VP treatments 1 and 2 (Figure 4).

Result 5: The average number of allocated punishment points is smaller in VP treatments 1 and 2 than in VP treatment 3 and in the FP treatment.

All four pairwise comparisons between VP treatments 1 or 2 on one hand and VP treatment 3 or the FP treatment on the other hand reveal differences that are

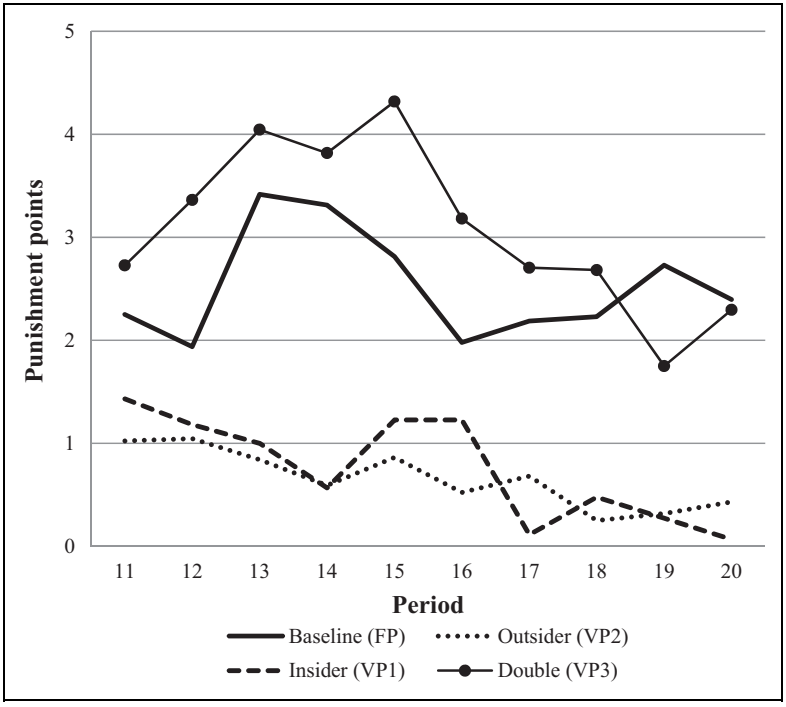

Figure 4. Average allocation of punishment points by treatment and period.

statistically significant at conventional levels ($p_{VP1,VP3} = .008$; $p_{VP1,FP} = .015$; $p_{VP2,VP3} = .006$; $p_{VP2,FP} = .012$). In contrast, no statistically significant difference exists between VP treatments 1 and 2 ($p_{VP1,VP2} = .870$) or between VP treatment 3 and the FP treatment ($p_{VP3,FP} = .559$).

Allocating punishment points is costly in our design; hence, a rational and purely self-regarding subject will *never* allocate punishment points. However, our results suggest that subjects' motivation for allocating costly punishment points does have an instrumentally rational component: subjects' willingness to allocate punishment points is stronger when no escape option is available (FP treatment and VP treatment 3) than when the presence of an escape option makes punishment futile (VP treatments 1 and 2).

Who are the punished subjects? Our results indicate that punishments are instrumental in the FP treatment and in VP treatment 3: here, the more a subject contributes, the less its payoff is reduced by punishment (see Figure S1 in the supplementary material). No similar pattern exists in VP treatments 1 or 2. In VP treatment 1, the average punishment of subjects that contribute little (0–5 ECUs) is rather mild, perhaps because punishing subjects understand that harsher punishments would cause reduced

participation. In VP treatment 2, average punishments are very low, presumably because subjects participate to avoid punishment.

Our results do not reveal any discernible pattern concerning which subjects choose to punish (see Figure S2 in the supplementary material). In VP treatment 1, subjects that contribute more to the public good also tend to sanction low contributions more severely. However, in the FP treatment, allocated punishments are roughly independent of own contribution level. And in VP treatments 2 and 3, subjects with contributions in the 11 to 15 ECUs range tend to impose the most severe punishments.

Conclusion

Our experimental results add to existing knowledge both in the MEA literature debating the effect of compliance enforcement and in the experimental literature on public goods provision. Concerning the MEA literature, they suggest that the managerial and enforcement schools are *both* right, but under different circumstances. Without participation enforcement, compliance enforcement *fails* to enhance compliance. However, with participation enforcement, it enhances compliance *substantially*. As most existing MEAs lack participation enforcement, designing institutions for compliance enforcement in these MEAs may well be a waste of time, as the managerial school argues. The reason is that countries reluctant to contribute to problem solving will unlikely participate in the first place. However, designing effective MEAs for particularly challenging problems such as climate change may be difficult without the use of incentives for inducing reluctant countries to participate. For MEAs addressing such problems, institutions for compliance enforcement may therefore be essential, as the enforcement school argues.

More generally, our results indicate that enforcing only compliance or only participation will have little or even no effect on an MEA's public goods provision. Compliance enforcement without participation enforcement will cause free riding to take the form of nonparticipation, whereas participation enforcement without compliance enforcement will cause free riding to take the form of noncompliance. To be effective, MEAs aiming to provide a public good must deter both types of free riding; thus, they must enforce both participation and compliance.

Concerning the experimental literature on public goods provision, our experimental results establish that the presence of both free-rider options can, depending on the enforcement system, significantly hamper public goods provision. In particular, they show that previous experimental findings showing that enforcement tends to boost cooperation are valid only in settings with forced (or enforced) participation. They also reveal that subjects' willingness to allocate costly punishment points is significantly stronger when the enforcement system permits punishment of both types of free riding than when it permits punishment of only one type.

Our results help resolve an apparent tension between part of the MEA literature (specifically, that concerning the managerial school's perspective on compliance

enforcement) and the experimental literature. While the managerial school considers compliance enforcement largely pointless, the experimental literature finds that compliance enforcement enhances public goods provision substantially. Our experiment shows that these divergent findings are not only compatible, but also simply what one should expect. Managerial-school scholars have largely studied compliance in MEAs without participation enforcement; hence, it is only natural that they find no effect of compliance enforcement. In contrast, experimental scholars have mostly considered settings with FP; hence, it is also only natural that they find a substantial effect of enforcement.

Finally, our results also help clarify the potential relevance of the experimental literature for the study of MEAs. As most of this literature concerns experiments that implement FP, those experiments' results are relevant primarily for the study of those (relatively few) MEAs that have participation enforcement. Thus, for MEAs *without* participation enforcement, the results from the experimental literature must be applied with particular care.

We end by suggesting two alternative designs that might come even closer to a real-world MEA setting. It would be interesting to see if our main results hold up even under these alternative designs. First, in our experiment, subjects did not make any (nonbinding) commitments concerning their contribution. Requiring subjects choosing to participate in the project to do so would arguably increase the experiment's resemblance to real-world MEAs, where such commitments are indeed common (e.g., in the form of targets for emissions reductions).

Second, in our experiment, subjects decided in every round whether to participate in the project (and then those who chose to participate made a second decision concerning their contribution's size). In real-world MEAs, countries rarely (if ever) move repeatedly in and out, although participating countries may be free to withdraw (e.g., Canada withdrew from Kyoto in 2011) and countries that did not participate initially may be free to join later (e.g., Australia ratified Kyoto two years after its entry into force). An alternative design might let subjects make a onetime decision of whether to participate and then only make repeated contribution decisions. Alternatively, subjects might be permitted to change their initial participation decision *once* (rather than in every period).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Hovi acknowledges the funding from the Research Council of Norway through grant 209701/E20 for CICEP. Aakre and Hovi acknowledge the funding from the Research Council of Norway through grant 185508.

Supplemental Materials

The online [appendices/data supplements/etc.] are available at http://jcr.sagepub.com/supplemental.

Notes

1. Much can be said for including positive as well as negative incentives in the definition of enforcement (Breitmeier, Young, and Zürn 2006, 148-49). Positive incentives for compliance and participation may include side payments, issue linkages (including trade restrictions), and allocation of entitlements such as emission permits (Barrett and Stavins 2003, 360).
2. "Reciprocal measures" refers to conditional material punishments (Axelrod and Keohane 1986) and should not be confused with "reciprocity" in the sense of conditional social preferences.
3. Some scholars distinguish three or more explanatory models of compliance; see Underdal (1998) and Breitmeier, Young, and Zürn (2006).
4. To be fair, both the managerial school and the enforcement school do consider participation to some extent. For example, Chayes and Chayes (1995) favor managerial strategies that permit countries to ignore certain obligations (i.e., not to participate in certain parts of the agreement) while capacity-building efforts are being made. Similarly, Downs, Rocke, and Barsoom (1996) emphasize that the contents of and participation in multilateral environmental agreements (MEAs) are endogenous to countries' underlying interests. However, neither school considers the interaction between compliance enforcement and participation enforcement in the way we do here.
5. However, MEAs are not "complete contracts" that unambiguously prescribe an action for each participating country in every conceivable state of the world. Thus, even in MEAs with enforcement signatories must exert some discretion in assessing what constitutes acceptable behavior and what constitute *serious enough* transgressions to warrant punishment. Our design allows for such assessments.
6. Worth noting, however, is that many MEAs include a minimum participation clause stating that entry into force requires a certain threshold level of participation. Studies analyzing the effect of such a clause on MEA participation find that a credible minimum participation clause may induce countries to become signatories (e.g., Carraro, Marchiori, and Orrefice 2009); however, because some countries may be better off by cooperating even if the threshold is not reached, a high minimum participation threshold may not be credible (Barrett 2003).
7. For a comprehensive review of the literature on international compliance, see Raustiala and Slaughter (2002).
8. This is not to say that noncompliance with MEAs is nonexistent. For example, as many as ten (of twenty-five) parties to the 1999 Gothenburg Protocol under the Convention on Long-Range Transboundary Air Pollution failed to meet at least one of their four emissions reduction targets (sulfur dioxide, nitrogen oxides, volatile organic compounds, ammonia) by the 2010 deadline.

9. While Downs, Rocke, and Barsoom (1996) consider *international* enforcement, Dai (2005, 2007) suggests that countries comply because of *domestic* sources of enforcement, such as the electoral leverage and the informational status of domestic constituencies.
10. International law requires states to comply with agreements to which they choose to be a party, but it does not require them to become a party in the first place (Barrett and Stavins 2003).
11. For a similar interpretation, see Victor (1998).
12. Isaac, McCue, and Plott (1985) initiated this literature. Ledyard (1995), Fehr and Schmidt (1999), and Plott and Smith (2008, part 6) provide excellent reviews.
13. Ostrom, Walker, and Gardner (1992) introduced this extension.
14. A challenge for this conjecture is that restarting the experiment with the same subject group tends to reproduce the same pattern (Andreoni and Croson 2008).
15. Ostrom (2000, 141) denotes such subjects "willing punishers," stating that a willing punisher "will expend personal resources to punish those who make below-average contributions to a collective benefit, including in the last period of a finitely repeated game."
16. Participation in Kyoto's second commitment period, running from 2013 to 2020, is even more limited; Belarus, Canada, Japan, New Zealand, Russia, and Ukraine have all declined to participate.
17. Other MEAs that may be placed in this category include MARPOL and CITES. MARPOL bars tankers that violate appropriate equipment standards from doing business in signatory ports, regardless of whether they carry a signatory flag. Similarly, under CITES (article X) trade with a nonparty requires that this nonparty provides documentation comparable to that required of a party and violations are enforceable through so-called CITES implementation legislation, that is, domestic laws permitting each party to implement and enforce CITES regulations.
18. This assumption ensures that a unique subgame-perfect equilibrium exists and reflects the intuition that a country cannot join without making at least a small contribution toward reaching the MEA's goal. Moreover, being part of international negotiations may entail political and transaction costs.
19. Decisions are public knowledge in the sense that all subjects are informed that all subjects can observe the number of participating group members and each participant's contribution. However, participants' identities remain anonymous.
20. Notice that the three voluntary participation (VP) treatments are identical in the first ten periods.
21. In the baseline treatment, the subgame-perfect equilibrium is that no player ever contributes and no punishment is imposed.
22. In the unlikely—though logically possible—event of a subject's going bankrupt, the administrator could allocate credits to the bankrupt subject (to prevent termination of the session). We did not inform subjects about this credit option, and we never used it, as no subject went bankrupt in our experiment.
23. Notice that in the forced participation (FP) treatment, all group members were necessarily insiders.

24. The alternative design of rematching groups after each period tends to counteract implicit coordination and has been shown to reduce contribution levels under FP; however, the dynamics of contributions tend to be similar to those obtained with our matching protocol (Fehr and Schmidt 1999).
25. We are indebted to Simon Gächter for his immediate generosity in sharing his *z*-tree files and instructions with us. Our instructions and *z*-tree files are, of course, available upon request.
26. Average hourly earnings correspond to the going hourly rate for student research assistants at the second author's institution.
27. Our data contain time trends not accounted for by our averages (but are clearly visible in the figures). For formal tests of these time trends, see the supplementary material.
28. If only one subject were to participate, the insider punishment treatment would be stable in terms of participation, since this insider would not be able to punish itself (and presumably would not do so even if it could).
29. Recall that in the FP treatment all subjects are insiders.
30. The decline in contributions from periods 1 through 10 to periods 11 through 20 in VP treatment 1 may be part of a general time trend. To determine whether it is, one must run a twenty-round baseline treatment without enforcement. The conclusion we can legitimately draw with our design is that some enforcement regimes (VP3 and FP) enhance contributions substantially, while others (VP1 and VP2) enhance contributions far less or not at all.
31. $p_{VP1,VP2} = .793$, $p_{VP1,VP3} = .393$, $p_{VP2,VP3} = .622$, $p_{VP1,FP} = .006$, $p_{VP2,FP} = .012$, $p_{VP3,FP} = .002$.
32. $p_{VP1,VP2} = .017$, $p_{VP1,VP3} = .002$, $p_{VP2,VP3} = .008$, $p_{VP1,FP} = .000$, $p_{VP2,FP} = .000$, $p_{VP3,FP} = .295$.
33. T values: VP1 versus VP2 $= -0.02$; VP2 versus VP3 $= 0.39$; VP1 versus VP3 $= 0.43$.
34. T values: VP1 versus VP2 $= 1.25$; VP2 versus VP3 $= -2.37$; VP1 versus VP3 $= -3.44$.
35. We also ran our regressions on individual contributions (available from the authors upon request; $N = 1{,}800$), with standard errors clustered on unique groups, and with random group effects. Our results are robust to these alternative specifications as well.

References

Ahn, Toh-Kyeong, R. Mark Isaac, and Timothy C. Salmon. 2009. "Coming and Going: Experiments on Endogenous Group Sizes for Excludable Public Goods." *Journal of Public Economics* 93 (1): 336-51.

Andreoni, James, and Rachel Croson. 2008. "Partners Versus Strangers: Random Rematching in Public Goods Experiments." In *Handbook of Experimental Economic Results. Volume 1*, edited by Charles R. Plott and Vernon L. Smith, 776-83. Amsterdam: North Holland.

Axelrod, Robert, and Robert O. Keohane. 1986. "Achieving Cooperation under Anarchy: Strategies and Institutions." In *Cooperation under Anarchy*, edited by Kenneth A. Oye, 226-54. Princeton, NJ: Princeton University Press.

Barrett, Scott. 2003. *Environment and Statecraft: The Strategy of Environmental Treaty-making*. Oxford, UK: Oxford University Press.

Barrett, Scott, and Robert Stavins. 2003. "Increasing Compliance and Participation in International Climate Change Agreements." *International Environmental Agreements: Politics, Law and Economics* 3 (4): 349-76.

Brack, Duncan. 2003. "Monitoring the Montreal Protocol." In *Verification Yearbook 2003*, edited by Trevor Findlay, 209-26. London: VERTIC.

Bratberg, Espen, Sigve Tjøtta, and Torgeir Øines. 2005. "Do Voluntary International Environmental Agreements Work?" *Journal of Environmental Economics and Management* 50 (3): 583-97.

Breitmeier, Helmut, Oran R. Young, and Michael Zürn. 2006. *Analyzing International Environmental Regimes: From Case Study to Data Base*. Cambridge, MA: MIT Press.

Brown Weiss, Edith, and Harold K. Jacobson, eds. 1998. *Engaging Countries: Strengthening Compliance with International Environmental Accords*. Cambridge, MA: MIT Press.

Carraro, Carlo, Carmen Marchiori, and Sonia Orrefice. 2009. "Endogenous Minimum Participation in International Environmental Treaties." *Environmental and Resource Economics* 42 (3): 411-25.

Charness, Gary B., and Chun-Lei Yang. 2010. "Endogenous Group Formation and Public Goods Provision: Exclusion, Exit, Mergers, and Redemption." Working Paper, Department of Economics, University of California, Santa Barbara.

Chayes, Abram, and Antonia H. Chayes. 1993. "On Compliance." *International Organization* 47:175-205.

Chayes, Abram, and Antonia H. Chayes. 1995. *The New Sovereignty*. Cambridge, MA: Harvard University Press.

Cherry, Todd, and David M. McEvoy. 2013. "Enforcing Compliance with Environmental Agreements in the Absence of Strong Institutions: An Experimental Analysis." *Environmental and Resource Economics* 54 (1): 63-77.

Dai, Xinyuan. 2005. "Why Comply? The Domestic Constituency Mechanism." *International Organization* 59 (2): 363-98.

Dai, Xinyuan. 2007. *International Institutions and National Policies*. Cambridge: Cambridge University Press.

Dannenberg, Astrid. 2012. "Coalition Formation and Voting in Public Goods Games." *Strategic Behavior and the Environment* 2 (1): 83-105.

Dannenberg, Astrid, Andreas Lange, and Bodo Sturm. 2014. "Participation and Commitment in Voluntary Coalitions to Provide Public Goods." *Economica* 81 (322): 257-75.

Downs, George W., David M. Rocke, and Peter N. Barsoom. 1996. "Is the Good News about Compliance Good News about Cooperation?" *International Organization* 50:379-406.

Fehr, Ernst, and Urs Fischbacher. 2005. "The Economics of Strong Reciprocity." In *Moral Sentiments and Material Interests. The Foundations of Cooperation in Economic Life*, edited by Herbert Gintis, Samuel Bowles, Robert T. Boyd, and Ernst Fehr, 151-91. Cambridge, MA: MIT Press.

Fehr, Ernst, Urs Fischbacher, and Simon Gächter. 2002. "Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms." *Human Nature* 13:1-25.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *The American Economic Review* 90 (4): 980-94.

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415:137-40.
Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics* 114 (3): 817-68.
Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rochenbach. 2006. "The Competitive Advantage of Sanctioning Institutions." *Science* 312 (5770): 108-11.
Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment across Societies." *Science* 319 (5868): 1362-67.
Isaac, R. Mark, Kenneth F. McCue, and Charles R. Plott. 1985. "Public Good Provision in an Experimental Environment." *Journal of Public Economics* 26:653-70.
Kosfeld, Michael, Akira Okada, and Arno Riedl. 2009. "Institution Formation in Public Goods Games." *The American Economic Review* 99 (4): 1335-55.
Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *The Handbook of Experimental Economics*, edited by John Kagel and Alvin Roth, 111-94. Princeton, NJ: Princeton University Press.
McDermott, Rose. 2011. "New Directions for Experimental Work in International Relations." *International Studies Quarterly* 55 (2): 503-20.
McEvoy, David M. 2010. "Not it: Opting out of Voluntary Coalitions that Provide a Public Good." *Public Choice* 142 (1): 9-23.
McEvoy, David M., James J. Murphy, John M. Spraggon, and John K. Stranlund. 2011. "The Problem of Maintaining Compliance within Stable Coalitions: Experimental Evidence." *Oxford Economic Papers* 63:475-98.
Ostrom, Elinor. 2000. "Collective Action and the Evolution of Social Norms." *Journal of Economic Perspectives* 14 (3): 137-58.
Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and without a Sword: Self-Governance Is Possible." *American Political Science Review* 86 (2): 404-17.
Oye, Kenneth. 1985. "Explaining Cooperation under Anarchy: Hypotheses and Strategies." *World Politics* 38 (1): 1-24.
Plott, Charles R., and Vernon L. Smith, eds. 2008. *Handbook of Experimental Economic Results*, vol. 1. Amsterdam: North Holland.
Raustiala, Kal, and Anne Marie Slaughter. 2002. "International Law, International Relations, and Compliance." In *Handbook of International Relations*, edited by Walter Caerlsnaes, Thomas Risse, and Beth A. Simmons, 538-58. London: Sage.
Raustiala, Kal, and David G. Victor. 1998. "Conclusions." In *The Implementation and Effectiveness of International Environmental Commitments: Theory and Evidence*, edited by David G. Victor, Kal Raustiala, and Eugene B. Skolnikoff, 659-708. Cambridge, MA: MIT Press.
Ringquist, Evan J., and Tatiana Kostadinova. 2005. "Assessing the Effectiveness of International Environmental Agreements: The Case of the 1985 Helsinki Protocol." *American Journal of Political Science* 49 (1): 86-102.
Tallberg, Jonas. 2002. "Paths to Compliance: Enforcement, Management, and the European Union." *International Organization* 56 (3): 609-43.
Tingley, Dustin H. 2011. "The Dark Side of the Future: An Experimental Test of Commitment Problems in Bargaining." *International Studies Quarterly* 55 (2): 521-44.

Underdal, Arild. 1998. ‘‘Explaining Compliance and Defection: Three Models.’’ *European Journal of International Relations* 4 (1): 5-30.

Urpelainen, Johannes. 2011. ‘‘The Enforcement-exploitation Trade-off in International Cooperation between Weak and Powerful States.’’ *European Journal of International Relations* 17 (4): 631-53.

Victor, David G. 1998. ‘‘The Operation and Effectiveness of the Montreal Protocol’s Non-compliance Procedure.’’ In *The Implementation and Effectiveness of International Environmental Commitments: Theory and Evidence*, edited by David G. Victor, Kal Raustiala, and Eugene B. Skolnikoff, 137-76. Cambridge, MA: MIT Press.