Appalachian
STATE UNIVERSITY

Department of Economics Working Paper

Number 16-09 | August 2016

Inequality Aversion and Coalition Formation

David M. McEvoy
Appalachian State University

John K. Stranlund
University of Massachusetts Amherst

Department of Economics
Appalachian State University
Boone, NC 28608
Phone: (828) 262-2148
Fax: (828) 262-6105
www.business.appstate.edu/economics

Inequality Aversion and Coalition Formation

David M. McEvoy (corresponding author)
Department of Economics
Appalachian State University
Boone, NC 28692
Email: mcevoydm@appstate.edu
Phone: 828-262-6126

John K. Stranlund
Department of Resource Economics
University of Massachusetts Amherst
Amherst, MA 01003

Abstract: We explore the formation of coalitions to provide a public good when some players are averse to payoff inequality between coalition members and non-members. A model is presented to demonstrate how inequality-averse preferences could cause players to deliberately block profitable but inequitable coalitions from forming, and how the likelihood of such blocks is affected by the magnitude of payoff inequality. We then empirically examine coalition formation rates using laboratory experiments. Our results show that profitable coalitions are less likely to form the bigger the gap in payoffs between members and freeriding non-members. The experimental design allows us to tease out potentially confounding effects between the level of inequality and the minimum number of players required to make the coalition profitable. As predicted, controlling for the size of the participation threshold, we find that coalition formation rates fall as the payoff gap between members and non-members is increased.

Keywords: self-enforcing agreements; inequality aversion; coalitions; experiments; public goods

1. Introduction

When resource users cannot rely on a regulatory body to adequately provide public goods or manage common resources they must coordinate their decisions endogenously by creating new institutions. Important examples include the provision of global public goods like nuclear non-proliferation and climate change mitigation. Other examples include local resource management in the absence of strong or effective regulatory institutions (e.g., fishing communities in developing countries). Endogenous institution formation typically involves individuals voluntarily agreeing to forfeit some of their sovereignty in order to reach a collective goal. In this paper we examine voluntary formation of coalitions to provide public goods in which members agree to maximize their collective earnings. Such coalitions have been explored in depth in the literature (e.g., Barrett 1994; Rubio and Ulph 2006; Kolstad 2007; McGinty 2007; Eichner and Pethig 2013; Finus and Pintassilgo 2013). The equilibrium size of such coalitions (often called *self-enforcing* coalitions) is the smallest profitable coalition; that is, the smallest coalition required for its members to be materially better off within the coalition than they would be in the absence of any coalition forming. Non-members, if any, are even better off than coalition members because they get to freeride on the members' provision of the public good. The difference between member and non-member payoffs is the cost of providing the public good.

Two important points regarding equilibrium coalitions that provide public goods are well established in the literature (beginning with Barrett 1994). First, coalitions are largest when the cost of providing the public good is high relative to the benefits. This means that coalitions are large when there are relatively small gains from cooperation, and they are small when cooperation could yield large gains. This result, however paradoxical, is a pessimistic prediction for individuals trying to achieve cooperative goals. The second important point is that equilibrium coalitions are extremely fragile. Because an equilibrium coalition is the smallest possible coalition that is mutually beneficial to its members, they are always right at the edge of being made worse off—a single defection from an equilibrium coalition will make the remaining members worse off than if no coalition had formed. This may suggest that the formation of coalitions is sensitive to individual preferences over non-material earnings, such as those involving preferences for equity.

In this paper we explore how coalition formation rates are affected by individual preferences over payoff inequality. Inequality aversion is potentially important in the analysis of

coalition formation and endogenous institutions in general because of the gap in payoffs between cooperating members and free-riding non-members. If resource users only cared about their material wellbeing then the equilibrium coalition size is well established. However, if utility also depends on relative earnings, then other equilibrium coalition sizes are possible. In this paper we build inequality aversion into a standard theoretical model of coalition formation and public good provision. We use the model to demonstrate how coalition formation rates change with the payoff gap between coalition members and non-members when members may feel disutility from earning less than freeriders. Conceptually, an inequality-averse individual may block the formation of a coalition if her material gain from joining the coalition is less than the disutility she experiences from earning less than freeriders. Assuming that the likelihood that a coalition will be blocked is increasing in this disutility, we examine how coalition formation rates are affected by the payoff gap between coalition members and non-members. In the process we identify two countervailing effects on coalition formation rates of changing this payoff gap. The first is a direct payoff inequality effect which leads to more blocked coalitions as the payoff gap between members and non-members is increased. The second is a threshold effect that is due to the fact that increasing the payoff gap leads to larger minimally profitable coalitions, and thus higher participation thresholds for effective coalitions to form. This threshold effect motivates fewer blocked coalitions because there are then fewer freeriders. The combined effects of increasing the payoff gap on coalition formation rates are non-monotonic. The likelihood of coalition blocks increases with the payoff gap between coalition members and non-members for low payoff gaps, but decreases for high payoff gaps. The reason is that the payoff inequality effect dominates the threshold effect for low payoff gaps, while the threshold effect dominates for high payoff gaps.

We then test these predictions with a series of laboratory experiments in which subjects choose whether or not to join a coalition to provide a public good. Our experiments isolate the effects of payoff inequality, the threshold effect, and their combined effects on coalition formation rates. As predicted, controlling for the size of the participation threshold, we find that coalition formation rates fall as the payoff gap between members and non-members is increased. However, our results do not support the hypothesis that coalition formation rates increase with higher participation thresholds, holding the payoff gap constant. This may be due to a coordination problem of reaching a higher threshold that offsets the reduction in the disutility

from unequal payoffs that comes from having fewer freeriders. Previous experimental studies of threshold public goods games have found that increasing the minimum number of required contributors reduces the likelihood of reaching the threshold (Van de Kragt et al. 1983; Dawes et al. 1986).[1] Finally, given the parameters we chose for our experiments, we expected that the combined effects of payoff inequality and higher thresholds would lead to lower coalition formation rates. In fact, this is what we observe in our experiments.

A number of theoretical papers have explored the role of inequality on equilibrium coalition sizes. Lange and Vogt (2003) model international environmental agreements when countries have other regarding preferences. They explore agreement formation when players have preferences toward equity, reciprocity and competition leaning on the theory developed by Bolton and Ockenfels (2000). They find that if countries put significant weight on earning an equal payoff share, then other equilibria are possible compared to standard models of payoff maximizing agents. Preferences toward equity can increase coalition sizes and even the grand coalition is possible. Lange (2006) adapts the standard model of international environmental agreements to allow for inequality-averse preferences across different forms of inequity among heterogeneous countries. His model imposes payoff inequality exogenously by dividing the world into developed and developing countries. He also finds that inequality aversion can motivate larger coalition sizes. A more recent paper by Kolstad (2014) provides a review of the literature on international environmental agreements and social preferences, and introduces a model of agreement formation when countries have preferences according to the theory developed by Charness and Rabin (2002). The model specifies a utility function that includes preferences toward equity, altruism and efficiency. Kolstad (2014) finds that when countries have such preferences, equilibrium coalition sizes can be smaller than those predicted by pure self-interest. The reason is that altruism and preferences for efficiency increase a player's utility

[1] Van de Kragt et al. (1983) and Dawes et al. (1986) report results from public good games with a "minimum contributing set" (MCS), where the MCS is defined as the smallest number of contributors required for a public good to be provided. Players make binary, all-or-nothing decisions whether to contribute $5 to a public account or to keep the $5. Both studies examine groups of seven and compare results when the MCS is set to three or five. In both studies the percentage of times the public good is provided decreases when the MCS increases from three to five. Both studies also report no significant effect of including a money-back-guarantee (i.e., contributions are returned if the MCS is not met) on individual contributions. One notable difference between the threshold in these early studies and the one in ours is that in their design any contributions in excess of the MCS are wasted, where in our experiments the provision of the public good is linear in the number of contributors.

from cooperating (relative to only considering material payoffs) and so it is possible that fewer coalition members are needed to make cooperation desirable.

As others have found, our theoretical model suggests that inequality aversion can increase the size of equilibrium coalitions. However, our model also suggests that inequality aversion can cause coalitions that maximize collective payoffs to form less often than if all players had standard preferences. This insight is missing from the related literature.

The study closest to ours is Kosfeld et al. (2009). Like our work, they introduce a model of public good provision and coalition formation that allows for inequality-averse preferences according to the theory proposed by Fehr and Schmidt (1999). They also demonstrate that inequality aversion can increase equilibrium coalition sizes. In laboratory experiments they find that groups tend to form large coalitions, in fact, most often the grand coalition forms. Their approach, however, deviates in a fundamental way from the established literature on coalition formation. Their model and experiments do not specify the objective of successful coalitions *ex ante*. Rather, individuals join a coalition and then vote on what they will do *ex post*. Therefore, when players decide whether to join a coalition they are uncertain about what their payoffs will be. This is a striking difference from the deep literature on coalition formation, and is therefore unique. That Kosfeld et al. (2009) observe large coalition sizes in their experiments may be the product of inequality aversion or it could be the product of the uncertainty in members' payoffs at the time they decide to join a coalition. In contrast, our study uses a model firmly grounded in the literature on stable coalitions in which coalition members maximize their collective payoffs.

The rest of this paper proceeds as follows. In the next section we present our theoretical model of coalition formation to provide a public good with individuals who may have preferences over payoff inequality. In the third section, we present our experimental design and hypotheses derived from the theoretical model. We present the results of the experiments in the fourth section, and conclude in the final section.

2. The Model

Following Barrett (2003), Kolstad (2007) and McEvoy et al. (2011), consider n players with identical monetary payoffs who each make a binary decision whether to contribute to a public good. Player i's payoff function is

$$\pi_i = A + b(q_i + q_{-i}) - cq_i, \tag{1}$$

where $q_i = (0,1)$ is the decision whether to contribute, q_{-i} is the sum of contribution decisions made by all players other than *i*, $b > 0$ is the public marginal benefit from an individual contribution, $c > 0$ is the individual cost of contributing and *A* is a positive constant. Throughout, $c > b$ so no player contributes to the public good in a non-cooperative Nash equilibrium, and $nb > c$ so that all *n* players contribute to the public good in the social optimum. Thus, we have a familiar *n*-player Prisoners' Dilemma.

2.1 Coalition formation with payoff-maximizing players

Players have the opportunity to form a coalition to provide the public good. The game has two stages. In the first stage players decide independently and sequentially under perfect information whether to join a coalition. In the second stage, each player chooses whether to contribute to the public good. Those who join a coalition in the first stage make their contribution decisions to maximize their joint payoffs. If enough players join a coalition so that each is better off contributing to the public good than not contributing, all members contribute to the public good. A coalition whose members choose to contribute to the public good is called an *effective* coalition. All effective coalitions are profitable in the sense that each member is at least as well off contributing to the public good as they would be if no player contributed. If too few players join a coalition in the first stage to make contributing to the public good worthwhile, then none of them contribute and an effective coalition does not form. Players who choose not to join a coalition in the first stage never contribute to the public good in the second stage.

If we let *s* be the number of coalition members from the first stage of the game and the coalition jointly provides the public good, a member's payoff is

$$\pi^m(s) = A + bs - c, \tag{2}$$

where the superscript *m* denotes membership in a coalition. A non-member's payoff in this situation is

$$\pi^{nm}(s) = A + bs, \tag{3}$$

where the superscript *nm* denotes non-membership. Note that non-members earn more than members because they avoid paying the cost of providing the public good. The effect on coalition formation when the payoff gap *c* varies is the focus of our work.

The equilibrium coalition size in the simultaneous-choice version of this game is called a self-enforcing coalition. A self-enforcing coalition is one that is internally and externally stable in the sense that no member wishes to leave and no non-member wishes to join (Barrett 1994). It is well-established in this context that the only internally and externally stable coalition size is the smallest profitable (or effective) coalition size, which we denote as s^*. To determine s^*, set (2) equal to A (the non-cooperative payoff level) and solve for s to find $s = c/b$. s^* is the smallest integer value greater than or equal to, c/b, that is,

$$s^* = \min\{s \mid s \geq c/b\}. \tag{4}$$

Note that s^* increases (weakly) in the cost of providing the public good and decreases in the benefit. Therefore, as explained by Barrett (1994), equilibrium coalitions are relatively large (small) when the benefits from providing the public good are relatively small (large). Because coalitions with fewer than s^* members will not form, we will refer to s^* as the *participation threshold* for an effective coalition to form.

While the equilibrium coalition size s^* is most often derived under the assumption that players make simultaneous decisions to join or not, it is also the subgame-perfect coalition size in a game of sequential decisions, which is a feature of our experiments. To see why, first define a *critical* player as a player whose decision not to join a coalition in the first stage prevents an effective coalition from forming. Note that a player is critical if $n - s^*$ players have already refused to join a coalition. A critical player who simply maximizes her financial payoff will always choose to join a coalition. The reason is that if she refuses to join then all potential coalitions are unprofitable (i.e., the participation threshold is not met), implying that an effective coalition will not form and all players earn their noncooperative payoff A. On the other hand, if she joins she cannot do worse than the noncooperative payoff, which would occur if another critical player refuses to join, but she will be better off if all remaining players (who are themselves critical) join to form an effective agreement. While all critical players will join an agreement, all non-critical players will refuse to join because they earn more by staying out of an effective agreement. Therefore, an effective coalition with s^* members will form in the first stage of the game. In terms of timing, the first $n - s^*$ players to choose will decide to stay out of a

coalition while the remaining s^* will join.[2] In the second stage, the members of the coalition will provide their units of the public good while the non-members will not contribute.

2.2 Coalition formation with inequality-averse preferences

In this section we modify the model to incorporate inequality aversion following Fehr and Schmidt (1999). A similar approach is used in Kosfeld et al. (2009) and McEvoy et al. (2015). While these authors have established that inequality aversion can cause larger coalitions to form, we focus here on how inequality averse individuals may *block* profitable coalitions from forming in the sense that they refuse to join a coalition when they are critical for its formation. An inequality averse individual may block a profitable coalition if the disutility they feel from earning less than freeriders is great enough. Assuming that coalitions are more likely to be blocked the greater is this disutility, we examine the likelihood that coalitions are blocked as a function of the payoff gap between coalition members and non-members.

Define the utility (including inequality aversion) of a member of an effective coalition with *s* members as

$$u_i^m(s) = \pi_i^m(s) - \frac{\alpha_i}{n-1}\sum_{j\neq i}\max\left(\pi_j(s) - \pi_i^m(s), 0\right) - \frac{\beta_i}{n-1}\sum_{j\neq i}\max\left(\pi_i^m(s) - \pi_j(s), 0\right), \quad (5)$$

where $\alpha_i \geq 0$ captures an individual's loss from disadvantageous inequality (i.e., disutility when other players earn more), and $\beta_i \geq 0$ captures the loss from advantageous inequality (i.e., disutility when other players earn less). The third term in (5) is zero because all members earn equivalent payoffs so that $\pi_i^m(s) - \pi_j^m(s) = 0,$ and non-members earn strictly higher payoffs than members so that $\pi_i^m(s) - \pi_j^{nm}(s) = -c$ and $\max(\pi_i^m(s) - \pi_j^{nm}(s), 0) = 0.$ The second term in (5) reduces to $(n-s)c,$ because $\pi_j^m(s) - \pi_i^m(s) = 0$ and $\pi_j^{nm} - \pi_i^m(s) = c$ is multiplied by the number of non-members. The first term in (5) is equation (2). Therefore, a member's utility can be written as

[2] See McEvoy (2010) for an experimental analysis on the timing of decisions in this game.

$$u_i^m(s) = A + bs - c - \frac{\alpha_i c(n-s)}{n-1}. \quad (6)$$

Note that $bs - c$ is the monetary payoff from joining a coalition with s members. The fourth term is a coalition member's disutility from earning less than non-members. We will examine this term more thoroughly shortly.

Now define the utility of a non-member of a coalition with s members as

$$u_i^{nm}(s) = \pi_i^{nm}(s) - \frac{\alpha_i}{n-1}\sum_{j\neq i}\max\left(\pi_j(s) - \pi_i^{nm}(s), 0\right) - \frac{\beta_i}{n-1}\sum_{j\neq i}\max\left(\pi_i^{nm}(s) - \pi_j(s), 0\right). \quad (7)$$

The second term in (7) equals zero because $\pi_j^{nm}(s) - \pi_i^{nm}(s) = -c < 0$ and $\pi_j^m(s) - \pi_i^{nm}(s) = 0.$ The third term in (7) is equal to $sc,$ because $\pi_i^{nm}(s) - \pi_j^{nm}(s) = 0$ and $\pi_i^{nm}(s) - \pi_j^m(s) = c$ is multiplied by the number of members. The first term in (7) is equation (3), and therefore a non-member's utility can be expressed as

$$u_i^{nm}(s) = A + bs - \frac{\beta_i sc}{n-1}. \quad (8)$$

In (8), bs is a freerider's gain if a coalition of size s forms to provide the public good. The third term is the disutility a freeriding non-member experiences because she earns more than coalition members. Like Fehr and Schmidt (1999) we assume $\alpha_i > \beta_i$ so that an individual feels less pain from advantageous inequality than disadvantageous inequality. This assumption preserves the freeriding incentive because

$$u_i^m(s) - u_i^{nm}(s) = \frac{c}{n-1}\left(s(\beta_i - \alpha_i) - \alpha_i n - (n-1)\right) < 0$$

for $\alpha_i > \beta_i$.

In the previous section in which players were only concerned about their individual payoffs, the equilibrium coalition is the smallest profitable coalition defined by (4). Here we use the more general utility measure from (6) to solve for the minimum *preferred* coalition size for an inequality-averse individual. Solving $u_i^m(s) = A$ for s yields this individual's minimum preferred coalition size as

$$\hat{s}_i = \min\left\{s \mid s \geq \bar{s}_i\right\} \text{where } \bar{s}_i = \frac{c(n-1) + \alpha_i cn}{b(n-1) + \alpha_i c}. \quad (9)$$

It is straightforward to show that $\overline{s}_i$ is strictly increasing in α_i and it is equal to c/b when $\alpha_i = 0$. Consequently, $\hat{s}_i = s^*$ for an individual that does not experience disutility from disadvantageous inequality and $\hat{s}_i \geq s^*$ for an individual who does feel such disutility.

Since $\hat{s}_i \geq s^*$, an inequality-averse individual may demand that coalitions form that are larger than the participation threshold for an effective coalition to form. Thus, aversion to disadvantageous inequality can lead to the formation of larger coalitions depending on the prevalence of this preference in a group of players. However, this sort of inequality aversion can also lead to fewer coalitions forming. Individuals who do not feel disutility from disadvantageous inequality would always rather join a minimum profitable coalition than to see it fail. However, an individual who is averse to disadvantageous inequality may actually block a profitable coalition from forming.

To explore this possibility more thoroughly, return to the payoff of an inequality averse member of a coalition given by (6). Note that an individual will block a coalition of size *s* if (6) is strictly less than the noncooperative payoff *A*. Note also that an individual can only block a coalition if her joining is critical for its formation. This implies further that an individual can only block the minimum profitable coalition given by (4). Therefore, an inequality-averse individual blocks the minimum size profitable coalition if and only if

$$u_i^m(s^*) - A = bs^* - c - \frac{\alpha_i c(n - s^*)}{n-1} < 0, \;\; s^* = \min\{s \mid s \geq c/b\}.$$

The monetary payoff from joining a coalition with s^* members is $bs^* - c$. This is greater than or equal to zero, so a payoff-maximizing individual will never block a coalition that reaches the participation threshold. However,

$$\alpha_i c(n - s^*)/(n-1), \tag{10}$$

is the disutility an inequality-averse member feels because non-members earn more. It is clear that even though a coalition with s^* members is profitable, an inequality-averse individual may block the coalition if the disutility from earning less than freeriders is high enough. It is reasonable for us to assume that the higher is (10), then the more likely it is that a profitable coalition will be blocked by an inequality-averse individual.

To examine how the disutility from disadvantageous inequality changes with the payoff gap between coalition members and non-members, let the participation threshold be

$$s^* = \frac{c}{b} + \gamma,$$

where γ is a non-negative parameter used here to roughly account for the discreteness of coalition sizes. Plug this into (10), differentiate with respect to *c* and rearrange terms to obtain

$$\frac{\alpha_i}{b(n-1)}\left(\frac{b(n-\gamma)}{2} - c\right).$$

This indicates that the individual's disutility from disadvantageous inequality increases for *c* up to approximately $bn/2$ and then falls from that point to bn. Thus, increasing *c* (and the accompanying participation threshold) can increase the number of blocked coalitions for low initial levels of *c* and smaller thresholds, but decrease the number of blocked coalitions for higher initial levels of *c* and larger thresholds.

The intuition behind the non-monotonic effect of increasing the payoff gap between members and non-members on coalition formation rates is the following. From (10), note that the disutility from earning less than freeriders increases with the payoff gap, holding s^* constant. We refer to this effect as the *payoff inequality effect*. On the other hand, increasing s^* while holding *c* constant reduces the disutility from disadvantageous inequality because the number of freeriders is smaller. We refer to this countervailing effect as the *threshold effect*. In total, increasing the payoff gap between members and non-members has a non-monotonic effect on the disutility from earning less than freeriders, because the payoff inequality and threshold effects work in opposite directions. Moreover, the inequality effect dominates for low initial payoff gaps and accompanying participation thresholds, while the threshold effect dominates for higher initial payoff gaps and larger participation thresholds.

In the next section we present experiments designed to isolate the payoff inequality effect and the threshold effect of increasing the payoff gap between coalition members and non-members. If inequality aversion plays a significant role in coalition formation, we expect to observe a lower coalition formation rate as the payoff gap is increased, holding the participation threshold constant, but a higher formation rate as we increase the participation threshold, holding the payoff gap constant. What happens to the coalition formation rate as we allow both the payoff gap and the participation threshold to adjust depends on the parameters chosen for the experiment. We chose parameters that imply a lower coalition formation rate as the payoff gap and participation thresholds increase together.

3. Experimental Design

In each experimental treatment players decide individually whether or not they want to join a coalition. After all players decide, coalition members automatically make their binary contribution decisions to maximize joint payoffs. Therefore, if enough players join to make the agreement profitable, then an effective coalition forms and all coalition members contribute to the public good. If too few players join to reach the participation threshold then no one contributes. Non-members maximize their individual payoffs by not contributing whether an effective coalition forms or not. In our experiment all second-stage contribution decisions were automated and the single decision is whether or not to join the coalition.[3]

Players were placed in groups of 10 and had payoff functions matching equation (1). The first two treatments differ in the cost of contributing to the public good and the participation threshold for an effective coalition to form. In these treatments $A = 8$, $b = 3$ and c, the cost of contributing, was either 7 (low cost) or 15.10 (high cost). In the treatment in which $c = 7$ (labeled **T1**), by equation (4) the participation threshold for an effective coalition is three members (the smallest integer greater than $c/b = 7/3 = 2.5$). Players were informed of this threshold before making their decisions and their potential payoffs for all possible decisions were displayed in tabular form (**instructions available as a reviewer's appendix**). This payoff table is reproduced in Table 1 below. For this treatment, when three players join the coalition each earns \$10 while non-members earn \$17 (i.e., the payoff gap is \$7).[4] If fewer than three members join the coalition then no player contributes and each earns \$8. From the table is it clear that there is a financial incentive to be a member of an effective coalition (i.e., one with three or more players) relative to when a coalition fails (i.e., when less than three join). However, it is always more lucrative to be a non-member outside of an effective coalition.

Table 1: Payoff table for T1

# of OTHER players that join coalition	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

[3] There are a handful of experimental papers that explore coalition formation rates in public good games when members maximize their collective payoffs, including McEvoy (2010), McEvoy et al. (2011), Dannenberg (2012) and Dannenberg et al. (2014). These papers do not explicitly explore inequality-averse preferences.

[4] Earnings are reported in experimental dollars and ten experimental dollars exchanged for one US dollar. Subjects also earned an additional \$5 for showing up on time.

YOUR earnings if you **join**	8	8	10	13	16	19	22	25	28	31
YOUR earnings if you **don't join**	8	8	8	17	20	23	26	29	32	35

As the cost of providing the public good is increased to 15.10 in treatment **T2**, the threshold for an effective coalition increases to 6 players (the smallest integer greater than 15.10/3 = 5.03).[5] For a coalition of exactly six members each member earns \$11 and non-members each earn \$26. Thus, when the cost of public good provision is higher, the threshold coalition size is higher and so too is the gap between member and non-member earnings. Table 2 shows the payoffs for members and non-members for T2.

Table 2: Payoff table for T2

# of OTHER players that join coalition	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
YOUR earnings if you **join**	8	8	8	8	8	11	14	17	20	23
YOUR earnings if you **don't join**	8	8	8	8	8	8	26	29	32	35

Recall from our theoretical model that members of a coalition may feel disutility from earning less than freeriding non-members, and so may block profitable coalitions if this disutility is great enough. Moreover, increasing the payoff gap between members and non-members produces countervailing payoff inequality and threshold effects on the disutility from differential payoffs, and in turn on the likelihood that coalitions will be blocked. Using the parameters from T1 and T2, the values of the inequality disutility (10) are $\alpha_i(5.44)$ and $\alpha_i(6.67)$, respectively. The higher disutility from earning less than freeriders under T2 suggests the following hypothesis if there are inequality-averse subjects in our subject pool.

Hypothesis 1: Coalition formation rates are higher under T1 than under T2.

[5] The non-integer parameter value of 15.10 was chosen so that the minimum sized profitable coalition (from (4)) would be six players while also making sure that a member of a six-player coalition is financially better off compared to when no agreement forms. This is achieved because the difference between 6 and 5.03 is large enough to make membership rewarding (a member of a six-player coalition earns \$10.9 compared to \$8 without an agreement). To simplify the experiments the payoffs were rounded to integer values.

We conduct two more treatments to isolate the payoff inequality and threshold effects of increasing the payoff gap between coalition members and non-members from $7 to $15. In treatment **T3** the participation threshold is kept small at three members but we increase the payoff gap between members and non-members to mirror T2 at $15. The payoffs for T3 are found in Table 3.

Table 3: Payoff table for T3

# of OTHER players that join coalition	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
YOUR earnings if you **join**	8	8	11	14	17	20	23	26	29	32
YOUR earnings if you **don't join**	8	8	8	26	29	32	35	38	41	44

The final treatment **T4** maintains a high participation threshold at six members with a small payoff gap between members and non-members of $7. The payoffs for **T4** are found in Table 4.

Table 4: Payoff table for T4

# of OTHER players that join coalition	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
YOUR earnings if you **join**	8	8	8	8	8	10	13	16	19	22
YOUR earnings if you **don't join**	8	8	8	8	8	8	17	20	23	26

Given that the payoff inequality effect suggests that increasing the payoff gap between members and non-members can cause a coalition member to feel greater disutility from earning less than non-members, holding the participation threshold constant, we have the following hypothesis.

Hypothesis 2: Coalition formation rates are higher under T1 than under T3. For the same reason, coalition formation rates are higher under T4 than under T2.

Finally, because increasing the participation threshold is expected to reduce the disutility a coalition member feels from earning less than non-members, holding the payoff gap between members and non-members constant, we have our final hypothesis:

Hypothesis 3: Coalition formation rates are higher under T4 than under T1, and they are higher under T2 than under T3.

3.1 Protocols

For treatments **T1** and **T2** we utilize the experiments and data from McEvoy (2010) and McEvoy et al. (2011). These experiments were conducted at the University of Massachusetts Amherst using undergraduates recruited from the general student population. Treatments **T3** and **T4** use the same software program (programmed in Visual Basic) but with parameter values to reflect the payoffs in Tables 3 and 4. These experiments were conducted at Appalachian State University using undergraduates recruited from the general student population. Subjects made decisions individually and were given a set of instructions at their station that were read aloud by the moderator. Each session consisted of 20 subjects and participants were assigned randomly to one of the two groups of 10. Each period the groups were reshuffled so that the same ten people were never in the same group for more than one period. Two 13-period sessions were conducted for each of the four treatments yielding 520 individual observations and 52 group-level observations per treatment. Subjects were paid for all 13 periods. The frame of the experiment was kept neutral and players made decisions whether to join an *agreement* in which the members either jointly *produce* an unspecified product (if the participation threshold was met) or jointly *not produce* (if they did not meet the threshold).

As we are focused on the effect of payoff inequality on coalition formation rates we tried to mitigate coordination problems as much as possible. Therefore, the software was designed so that membership decisions were made sequentially with subjects having real-time information regarding the decisions made by the other nine members of the group. This information included the number of players that joined, the number that opted out and the number waiting to decide. Subjects were also updated with how many more players were needed to join to satisfy the participation threshold. The order of decision-making was endogenous (subjects could make decisions at any point in time), and once a subject made a decision the other players were notified and their decision buttons were made inactive for a fraction of a second so they could reevaluate their position. Once all players made their decision whether to join the agreement the period ended and the results were displayed.[6]

[6] In order to limit the length of time for each period, subjects had 60 seconds to make a decision. However, if someone made a decision in the last five seconds of the experiment and others were still undecided, an additional five seconds was added to the clock. If someone failed to decide before the time was up, she chose to not join the agreement by default. This feature was clearly explained to participants in their instructions.

4. Results

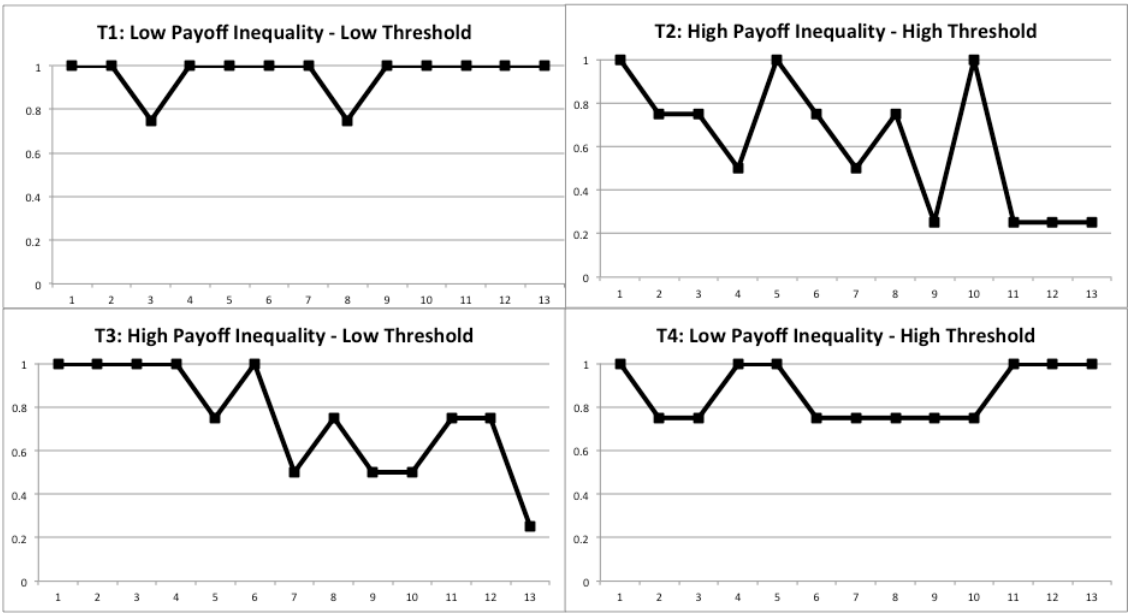
Table 5 presents group-level summary statistics on effective coalition formation rates, the average number of members when effective coalitions form and the average number of members for all coalitions (including failures) by treatment. Recall that if players are strict payoff-maximizers then coalitions are expected to form in 100% of trials for each of the four treatments. However, if at least some players have strong enough preferences for equal payoffs, then they may be better off causing a profitable coalition to fail rather than being a member. Our predictions about how coalition formation rates vary across treatments are contained in Hypotheses 1-3 in the previous section. Aggregating over all four treatments, profitable coalitions failed to form in roughly 20% of all trials. To illustrate changes over time, Figure 1 shows coalition formation rates over the 13 periods for each treatment. Each data point on the graph is the fraction of the four groups that formed effective coalitions. It is immediately clear that coalition formation rates are higher and more stable in the two treatments with lower payoff inequality.

Table 5: Effective coalition formation rates and average number of members

Treatment	Effective coalition formation rates	Avg. effective coalition size	Avg. coalition size overall
T1: low payoff inequality ($7)/ low threshold (3 members)	0.962 (0.027) [52]	4.26 (0.166) [50]	4.135 (0.184) [52]
T2: high payoff inequality ($15) /high threshold (6 members)	0.615 (0.068) [52]	6.188 (0.096) [32]	5.212 (0.206) [52]
T3: high payoff inequality ($15) /low threshold (3 members)	0.75 (0.061) [52]	3.231 (0.086) [39]	2.827 (0.128) [52]
T4: low payoff inequality ($7) /high threshold (6 members)	0.865 (0.048) [52]	6.311 (0.089) [45]	5.981 (0.147) [52]

Notes: The table presents group-level statistics, standard errors are in parentheses and sample sizes are in brackets.

Figure 1: Coalition formation rates by treatment over time



We supplement the summary statistics in Table 5 with a series of regression models of coalition formation to control for potential period and session effects and to rely on robust standard errors for our hypotheses tests. The dependent variable in each case takes on a value of one when an effective coalition forms. We regress this on dummy variables for the treatments (relative to T1), session and period.[7] Table 6 contains regression results for a linear probability, probit and logit specification (we dropped the coefficient estimates for periods and sessions in the table).

[7] The session dummies for each treatment are insignificant, but we reject the joint hypothesis ($p = 0.001$) that there are no period-level effects.

Table 6: Regression results on effective coalition formation

	Linear	**Probit**	**Logit**
T2	-0.385***	-1.607***	-2.965***
	(0.095)	(0.374)	(0.824)
T3	-0.269***	-1.175***	-2.244***
	(0.088)	(0.374)	(0.827)
T4	-0.154**	-0.728*	-1.412
	(0.076)	(0.400)	(0.869)
constant	1.202***	6.730***	19.071***
	(0.050)	(0.406)	(0.934)
model statistics	$n = 208$	n = 208	n = 208
	$F = 2.81$, $p = 0.000$	$\chi^2 = 1067.17$, $p = 0.000$	$\chi^2 = 1430.87$, $p =$
	$r^2 = 0.179$		0.000
hypotheses tests	T2 vs. T4, $p = 0.051$	T2 vs. T4, $p = 0.003$	T2 vs. T4, $p = 0.005$
	T3 vs. T2, $p = 0.360$	T3 vs. T2, $p = 0.103$	T3 vs. T2, $p = 0.105$
	T3 vs. T4, $p = 0.301$	T3 vs. T4, $p = 0.138$	T3 vs. T4, $p = 0.134$

Notes: In each model the dependent variable takes on a value of 1 when an effective coalition forms and 0 otherwise. The superscripts ***, ** and * indicate significance at the 0.01, 0.05 and 0.10 level, respectively.

4.1 Total effect of increasing the payoff gap between members and non-members (*Hypothesis 1*)

From Table 5 effective coalitions formed most often (96.2%) in T1, the scenario with low payoff inequality between members and non-members and a low participation threshold. Comparing this to the corresponding result from T2 (61.5%), the effective coalition rate decreased by about 35% in response to an increase in the cost of contributing to the public good and the accompanying increase in the participation threshold. This reduction is highly significant -

observe the coefficients for T2 in each of the regression models of Table 6. Given the parameters of our experiments, this effect of increasing the payoff is what we expected, thereby confirming Hypothesis 1.

4.2 Effect of increasing payoff inequality, controlling for the threshold effect (*Hypothesis* 2)
Now we keep the participation thresholds constant and examine the effects of an increase in payoff inequality on coalition formation rates, which we expect to be negative. Start by considering treatments T1 and T3 that have a participation threshold fixed at three members while the gap in payoffs jumps from $7 in T1 to $15 in T3. In fact, the formation rate is lower under T3 (75%) than under T1 (96.2%). This effect is captured by the coefficient on T3 in the regression models of Table 6, which are all highly significant. Now consider treatments T2 and T4 that each include a relatively high participation threshold of six members, but the payoff gap drops from $15 in T2 to $7 in T4. Again higher payoff inequality holding the participation threshold constant reduces the coalition formation rate (61.5% vs. 86.5% for T2 vs. T4). Hypotheses tests in Table 6 indicate that this effect is significant for each regression model. These results confirm Hypothesis 2 that increasing payoff inequality holding the participation threshold reduces coalition formation rates.

4.3 Effect of increasing participation threshold, holding the payoff gap constant (*Hypothesis 3*)
Contrary to Hypothesis 3, the threshold effects tend toward decreasing coalition formation rates, although these effects are not all statistically significant. Recall that the difference between T1 and T4 is that the payoff gap is constant at $7 while the participation threshold increases from three to six members. We expect that the coalition formation rate will be lower under T1, but in fact the formation rates are not very different (96.2% vs. 86.5% for T1 vs T4), and the hypothesis tests in Table 6 suggest that this difference is only weakly significant.[8] Similarly, we expect that coalition formation rates would be lower under T3 than under T2, but the opposite occurs (75% vs. 61.5% for T3 vs. T2). This effect is not significant at conventional levels for any of our regression specifications in Table 6, but the overall message is clear—we find no support for a positive threshold effect (Hypothesis 3).

[8] The coefficient estimate for T4 in Table 5 is insignificant with the logit specification, significant at the 10% level under a probit specification and significant at the 5% level in a linear model.

What could explain this result? One possibility is that a separate coordination effect (that is not modeled here) is at work that makes it harder for groups to form coalitions with higher participation thresholds. In fact, evidence from binary-choice threshold public goods games suggests that this may be a common finding (Van de Kragt et al. 1983; Dawes et al. 1986). Such a coordination effect could be working to offset the conceptual threshold effect in our experiments.

4.4 Coalition sizes

Recall that our theoretical model suggests that inequality-averse individuals will tend to require that larger coalitions form for their participation. If such preferences are prevalent in a group, then we might expect that the size of coalitions that form will exceed the participation threshold levels. From Table 5, we observe effective coalition sizes higher than the relevant threshold in all four treatments, but with substantial variability. The largest difference is found in T1 in which coalitions average 1.28 members more than the threshold of three. Recall that in this treatment both the payoff gap and the participation threshold are low. The next largest difference between the average coalition size and the threshold is observed in T4 in which payoff inequality remains low but the threshold is increased (0.311 members more than the threshold of six). In the two treatments in which payoff inequality is high (T2 and T3), coalition sizes are closest to the threshold levels. In summary, while we observe effective coalition sizes greater than the threshold in all four treatments, coalitions are relatively larger when the gap between member and non-member payoffs is smaller.

When considering the average size of all coalitions (including ones that did not reach the participation threshold) in Table 5, again, the only treatment with evidence of participation levels greater than the threshold is T1 (4.12 vs. 3). In contrast, we see lower participation rates compared to the threshold in T2 (5.212 vs. 6), T4 (5.981 vs. 6) and T3 (2.827 vs. 3).

4. Conclusion

Challenges like managing common resources, climate change mitigation, nuclear non-proliferation and the preservation of biological diversity require users to voluntarily restrict their

use of shared resources toward achieving a collective goal. In the absence of strong regulatory bodies, resource users typically coordinate their activities through the creation of new institutions in which the members commit to collectively managing their resources. Over the past two decades, theoretical research has explored the formation of such institutions through the concept of a *self-enforcing* coalition. The self-enforcing coalition is typically the smallest coalition required for its members to earn higher payoffs within a coalition than without one. Freeriding non-members, on the other hand, are always better off. Because the self-enforcing coalition is the smallest profitable coalition, its members are always right at the cusp of being made worse off. For this reason the formation of profitable coalitions may be sensitive to other-regarding preferences. In this paper we specifically explore how inequality-averse individuals can prevent profitable but inequitable coalitions from forming.

We have introduced a theoretical model of coalition formation to provide a public good when some players are averse to payoff inequality. We demonstrate that such players may block profitable coalitions because they would feel disutility from being in a coalition and earning less than freeriding non-members. Moreover, we identify two opposing effects of increasing the payoff gap between coalition members and non-members on the likelihood of coalitions forming. One is a direct payoff inequality effect while the other is an indirect participation threshold effect. The theoretical model suggests that increasing the payoff gap between members and non-members, holding the participation threshold constant, should lead to lower coalition formation rates, while increasing the threshold, holding the payoff gap constant, should lead to higher coalition formation rates. The overall effect of increasing the payoff gap depends on the parameters of the problem. We then test for these effects with a series of economic experiments, the results of which broadly support our hypotheses although not entirely. Given the parameters of our experiments, we expected that the overall effect of increasing the payoff gap between coalition members and non-members is to reduce the coalition formation rate, which is what occurs in our experiments. We also find strong support for the direct payoff gap effect. However, we do not observe higher coalition formation rates with higher participation thresholds, controlling for the payoff gap between coalition members and non-members. We suspect that there is an unspecified coordination effect at work here that offsets the theoretical threshold effect.

Our research contributes to the broad literature on endogenous institution formation to provide public goods. While others have shown theoretically that social preferences can affect the size of equilibrium coalitions, ours is the first empirical test of how varying the level of payoff inequality between members and non-members influences the likelihood of a profitable coalition forming. Although experiments necessarily oversimplify the relationship between resource users in a strategic environment, the method offers a way to evaluate the impact of specific policy components in a controlled environment. The experimental approach complements the theoretical analysis by testing its predictions and by highlighting the importance of behavioral elements that might not be present in the theory. Our results have implications for the design of effective institutions to provide public goods in the absence of a strong regulatory body. The findings suggest that profitability is a necessary but insufficient condition for coalitions to form to increase the provision of a public good beyond business as usual. We show that preferences toward equitable payoffs can prevent agreements from forming even when they are materially beneficial. In the context of climate change - where equity concerns are forefront in negotiations - the results suggest that agreements should be designed with such behavioral elements in mind. Given our findings, it may be wise to set minimum participation levels in agreements that reach beyond the profitability requirement in order to account for preferences for equity and fairness. In addition, although we do not model financial transfers in our analysis, our results highlight the important role they might play within voluntary institutions. Financial transfers could be utilized to satisfy both profitability and equity constraints in order to foster cooperation through coalition formation.

References

Bagnoli, M., & McKee, M. 1991. "Voluntary Contributions Games: Efficient Private Provision of Public Goods." *Economic Inquiry* 29: 351–366.

Barrett, S., 2003. *Environment and Statecraft: The Strategy of Environmental Treaty-Making*. Oxford and New York: Oxford University Press.

Barrett, S., 1994. "Self-Enforcing International Environmental Agreements." *Oxford Economic Papers.* 46(1): 878-94.

Bolton, G.E. and A. Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity and Competition. *American Economic Review* 90(1): 166-193.

Charness, G. and M. Rabin. 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics* 117(3): 817-869.

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

Dannenberg, A., A. Lange and B. Sturm. 2014. "Participation and Commitment in Voluntary Coalitions to Provide Public Goods." *Economica* 82(322): 257-275.

Dawes, R.M., J.M.Orbell, R.T. Simmons and A.J.C. van de Kragt. 1986. "Organizing Groups for Collective Action." *American Political Science Review* 80(4): 1171-1185.

Eichner, T. and R. Pethig. 2013. "Self-enforcing Environmental Agreements and International Trade." *Journal of Public Economics* 102(June): 37-50.

Fehr, E. and K. M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114(3): 817-68.

Finus, M. and P. Pintassilgo, 2013. "The Role of Uncertainty and Learning for the Success of International Climate Agreements." *Journal of Public Economics* 103(1): 29-43.

Kolstad, C., 2007. "Systematic Uncertainty in Self-Enforcing International Environmental Agreements." *Journal of Environmental Economics and Management.* 53(1): 68-79.

Kolstad, C. 2014. "International Environmental Agreements among Heterogeneous Countries with Social Preferences." *NBER Working Paper* No. 20204, June 2014.

Kosfeld, M., A. Okada and A. Riedl. 2009. "Institution Formation in Public Goods Games." *American Economic Review*, 99(4):1335-55.

Lange, A. and C. Vogt, 2003. "Cooperation in International Environmental Negotiations due to a Preference for Equity." *Journal of Public Economics* 87(9-10): 2049-67.

Lange, A., 2006. "The Impact of Equity-preferences on the Stability of International Environmental Agreements." *Environmental and Resource Economics* 34(2): 247-67.

McEvoy, D. and J. Stranlund, 2009. "Self-enforcing International Environmental Agreements with Costly Monitoring for Compliance." *Environmental and Resource Economics* 42():491-508.

McEvoy, D., 2010. "Not It: Opting Out of Voluntary Coalitions that Provide a Public Good." *Public Choice* 142(1): 9 – 23.

McEvoy, D., J. Murphy, J. Spraggon and J. Stranlund. 2011. "The Problem of Maintaining Compliance within Stable Coalitions: Experimental Evidence." *Oxford Economic Papers* 63(3): 475-498.

McEvoy, D., T. Cherry and J. Stranlund. 2015. "Endogenous Minimum Participation in International Agreements: An Experimental Analysis." *Environmental and Resource Economics* 62(4): 729-744.

McGinty, M., 2007. "International Environmental Agreements Among Asymmetric Nations." *Oxford Economic Papers* 59(1):45-62.

Rubio, S.J. and A. Ulph. 2006. "Self-enforcing International Environmental Agreements Revisited." *Oxford Economic Papers* 58(2): 233-263.

Van de Kragt, A.J.C., J.M. Orbell and R.M.Dawes. 1983. "The Minimal Contributing Set as a Solution to Public Goods Problems." *American Political Science Review* 77(1): 112-122.