NBER WORKING PAPER SERIES

INTERNATIONAL ENVIRONMENTAL AGREEMENTS AMONG HETEROGENEOUS COUNTRIES WITH SOCIAL PREFERENCES

Charles D. Kolstad

Working Paper 20204
http://www.nber.org/papers/w20204

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2014

Comments from Werner Güth, Kaj Thomsson and Philipp Wichardt and discussions with Gary Charness and Michael Finus have been appreciated. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

International Environmental Agreements among Heterogeneous Countries with Social Preferences
Charles D. Kolstad
NBER Working Paper No. 20204
June 2014
JEL No. H40,H41,Q5

ABSTRACT

Achieving efficiency for many global environmental problems requires voluntary cooperation among sovereign countries due to the public good nature of pollution abatement. The theory of international environmental agreements (IEAs) in economics seeks to understand how cooperation among countries on pollution abatement can be facilitated. However, why cooperation occurs when noncooperation appears to be individually rational has been an issue in economics for at least a half century. The problem is that theory suggests fairly low (even zero) levels of contribution to a public good and high levels of free riding. Experiments and empirical evidence with individuals suggests higher levels of cooperation. This is a major reason for the emergence in the 1990's and more recently of the literature on social preferences (also known as other-regarding preferences or prosociality) where participants account for their own well-being as well as that of others. This paper bridges the literature on cooperation among countries with the literature on cooperation among individuals. In particular, we introduce social preferences into a model of international environmental agreements. Focusing on Charness-Rabin social preferences, we find these preferences enlarge the set of conditions where cooperation is individually rational though such preferences also reduce the equilibrium size of a IEA for providing abatement. Although stable coalitions are smaller, more abatement may be provided by individual countries outside of a coalition structure. In contrast to much of the literature, we treat the size of agents as heterogeneous. Size of a country does not affect the incentives for forming a coalition but it does affect the aggregate level of abatement, suggesting that coalitions of large countries are more efficient than coalitions of small countries.

Charles D. Kolstad
Stanford Institute for Economic Policy Research
366 Galvez Street (Room 226)
Stanford, CA 94305-6015
and NBER
ckolstad@stanford.edu

I. INTRODUCTION

In the 1990's, a theoretical literature began to develop on the formation of coalitions to facilitate cooperation to solve externality problems (Barrett, 1994; Carraro and Siniscalco, 1993). Such coalitions are termed international environmental agreements (IEAs), much in the tradition of Olson's (1971) examination of groups to foster collective action. With some exceptions, that theoretical work has been pessimistic on the extent to which free-riding incentives can be overcome. This is discouraging for those hoping to solve major international environmental problems (such as climate change), solutions to which by their very nature demand cooperation.

The major reason that the standard static game theoretic model of an IEA suggests little gain from voluntary coalitions is that coalitions are held together purely by self-interest and self-interest usually drives participants to leave the coalition to join the fringe. One way the theoretical literature has dealt with this "problem" is to posit a repeated game structure with credible punishments for defection and non-cooperation. Drawing on the renegotiation-proof equilibria literature (Farrell and Maskin, 1989), Barrett (1999) and Asheim et al (2009) have demonstrated that in these conditions, more cooperation can be sustained than in the static case.

A parallel literature, largely disconnected from the IEA context, has focused on the behavior of individuals (not countries) in providing public goods. The literature on cooperation among individuals suggests that perhaps theory is too pessimistic on the extent of cooperation. Much of this literature involves laboratory experiments and calls into question the theoretical results that free riding is common and cooperation is difficult to sustain in the standard public goods problem (eg, Kim and Walker, 1984).

Partially in response to this disparity between empirics and theory, the notion of other-regarding or social preferences began to emerge in the early 1990's (though the terminology varies considerably), with Andreoni (1990) as one of the early contributors. The primary thrust of the social preference literature (eg, Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000;

Charness and Rabin, 2002) is that agents care about fairness as well as private payoffs and efficiency. Much of this literature is experimental, which some criticize as *ad hoc* and lacking an axiomatic foundation.

This paper provides a rigorous extension of the standard static IEA model to include social preferences, as developed in the public goods literature, reflecting a concern for self-interest, equity and efficiency.[1] Rather than offer our own theory of social preferences, we start with a particularly common form of social preferences, due to Charness and Rabin (2002). We then develop (1) a theory of voluntary contributions to public goods for the linear public goods game and (2) a theory of voluntary coalitions to provide public goods. Unlike many other papers, we allow heterogeneity with respect to country size, which yields significantly richer results.

A number of results with empirical consequence emerge from this paper. We find that theoretically, and in the standard linear public goods game, social preferences do enhance abatement and expand cooperation. Furthermore, participant size is not relevant to the decision to contribute, though it is relevant to the aggregate provision of abatement (larger countries contribute more). Although this is intuitive, the mechanism for expansion of cooperation is new.

II. BACKGROUND

In addition to the literature on international environmental agreements, a separate but important literature exists on voluntary contributions to public goods by individuals. Both are reviewed in the next two sections. It is important to recognize that the voluntary contributions literature focuses on individual consumers, each with a set of preferences. The literature on international environmental agreements focuses on decisions by countries, with each country treated as a welfare-maximizing entity. At a theoretical level, these two literatures are largely

[1] Lange and Vogt (2003) appears to be the first paper to introduce social preferences into IEAs, assuming homogeneous countries and using the preferences of Bolton and Ockenfels (2000), which focus on equity and self-interest. See also Lange (2006) and Dannenberg et al (2014).

interchangeable. However, experimental results applicable to voluntary contributions to public goods do not necessarily directly translate to public goods contributions by countries.

A. Private Provision of Public Goods

Inducing individual contributions to a public good in a noncooperative setting is a classic problem in public economics. Bergstrom et al (1986) provide the standard treatment of this problem, developing a simple model involving individual provision of a private good, x_i, a public good, g_i, and aggregate provision of the public good, G ($=\Sigma g_i$). Each identical agent (i) has simple preferences and an endowment of wealth, w_i, to be divided between x_i and g_i. The individual chooses x_i and g_i to maximize utility, subject to a budget constraint:

$$u(x_i,G) \text{ s.t. } x_i + g_i = w_i \qquad (1)$$

The first argument of u embodies the opportunity cost to the individual of providing the public good and the second term reflects the benefit of the aggregate provision. The authors show that in most cases there is a nonzero equilibrium provision of the public good. A second interesting result involves identical preferences but different wealth levels. In this case, there is a cutoff level of wealth. People who are poorer than the cutoff provide none of the public good whereas people above the cutoff provide a nonzero amount.

Andreoni (1988) uses this model to determine how contributions increase as the size of the economy (N —the number of individuals) increases. He shows that as N increases towards infinity, average individual contributions approach zero (though not the average *among the contributors*) and the size of the contributing group approaches zero. However, aggregate contributions approach a limit which is finite and nonzero. He points out that this result is at variance with casual empiricism that individuals do contribute to public goods, despite the economy being very large. For instance, according to Andreoni, half of all US households claim charitable donations on their tax returns (in the US, charitable donations are generally deductible from taxable income).

A significant body of experimental work has accumulated on this issue as well. Early experimental work on public good provision established that subjects tend to provide public goods at higher rates than predicted by the theory described above (Smith, 1980). Kim and Walker (1984) set up a laboratory experiment to test the "free rider hypothesis," which had been the subject of a number of papers in the 1970s (in the context of the prisoner's dilemma). The hypothesis simply is that individuals will prefer to free-ride rather than make contributions to the public good. The authors distinguish between "strong" free riders and other free riders. Strong free riders are closer to the theoretical behavior of contributing little to the public good. The authors show that although free-riding exists, they are not able to conclude that the free-riding is as strong as theory suggests. Isaac and Walker (1988) provide additional experimental evidence, exploring the role of an important variable, the *marginal per capita return* (MPCR). The MPCR is defined as the ratio of the marginal benefit to an individual of privately providing a public good to the marginal cost to the individual of that provision. Put differently, for every dollar a person spends on privately providing the public good, the MPCR measures how much the individual gets back. Clearly the MPCR is less than one (otherwise there is no issue). Higher MPCRs mean that the private gain from the public good is higher. A lower MPCR means that the individual is getting less private reward from providing the public good. Isaac and Walker (1988) demonstrate experimentally that MCPR is the primary determinant of contribution levels—there is no separate pure group size effect.[2] Furthermore, the authors demonstrate that the strong free rider effect is more pronounced for lower values of MPCR.

In an interesting review of this literature, Chaudhuri (2011) characterizes five main findings of the pre-1995 literature (attributing the last three to Ledyard, 1995): (1) in one shot versions of the noncooperative public goods game (described above) there is much less free-riding (more contribution) than predicted by theory; (2) if players repeat the one-shot game, free-riding increases with repeated interaction; (3) communication facilitates cooperation; (4)

[2] Isaac et al (1994) provide support for these findings using significantly larger groups.

thresholds facilitate cooperation; and (5) higher MPCRs lead to increased cooperation and decreased free-riding.

Over the past several decades, researchers have been moving beyond simple characterizations of payoffs to include a variety of "other-regarding preferences" or "social preferences" on the part of participants (see Sobel, 2005). One of the first extensions of this nature is the model of "impure altruism" by Andreoni (1989,1990), building in part on Olson (1971) and Becker (1994). Impure altruism holds that there are two avenues for personal utility gain from making a voluntary contribution to a public good: via the aggregate level of the public good and via a "warm glow" associated with the individual contribution. The individual may appear altruistic but that is because the individual obtains utility from giving. Thus the utility function in Eqn. (1) becomes $u(x_i,G,g_i)$. It is easy to see that including a private good dimension to public good contributions can remedy the apparent anomalies between the experimental results on free-riding and the theoretical results on contributions to public goods.

Other authors provide alternative models of contributing to public goods, always with the issue of free-riding as a motivator. Drawing on the fairness literature in psychology and economics (eg, Kahneman et al, 1986), Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) posit that inequality aversion drives cooperation. They propose the importance of inequality aversion as a dimension of utility that promotes cooperation and support the thesis with experimental evidence. Charness and Rabin (2002) present evidence in partial contradiction to this result, suggesting that efficiency also plays a role in outcomes in prisoner's dilemma games. To illustrate, in the Prisoner's Dilemma game shown in Fig. 1, theory would suggest defection will repeatedly occur. However, in an experimental setting, Charness et al (2008) find cooperation rates of 15%, 45% and 70% for values of x of 4, 5, and 6, respectively. This suggests more nuanced objectives. In particular, agents seem to be concerned with the total size of the "pie" as well as their own private payoffs.

Figure 1: Prisoner's Dilemma Payoffs from Charness et al (2008).

	B Cooperates	B Defects
A Cooperates	(x,x)	(1,7)
A Defects	(7,1)	(2,2)

Note: With payoff (a,b), a is payoff to player A and b is payoff to player B; $2<x<7$.

A number of authors have suggested that groups of people interacting strategically typically fall into at least two groups: self-interested and cooperators. This participant heterogeneity is said to explain the levels of cooperation observed in experimental and empirical data. Andreoni and Miller (2002) report experimental results for a dictator game and find 23% of respondents behave selfishly; the remainder show some degree of altruism. Fischbacher and Gachter (2010) conduct experiments with linear public goods game (similar to this paper) and find that 23% of participants are free-riders, contributing nothing; Fischbacher et al (2001) report a third are free-riders. Many of the remaining subjects are "conditional cooperators," cooperating conditional on others cooperating. Fehr and Schmidt (1999) review a number of papers with experimental results on public good games and conclude the papers, on average, involve a much higher fraction of pure free-riders (no contributions to the public good) – 73%. However, even with such a high fraction of free-riders, one can infer that the remaining subjects are behaving at variance with the purely selfish model. The upshot is that in groups, one might expect some heterogeneity in preferences—some agents are purely selfish and others display some sort of altruistic behavior.

Bliss and Nalebuff (1984), in a paper with a superb title, examine the case where individuals have different "abilities" or costs to supply the public good. They show that even in a noncooperative, repeated setting, the lowest cost individuals will eventually take it upon themselves to supply the public good.

B. International Environmental Coalitions/Agreements

The literature on international environmental agreements (IEA) focuses on countries as agents maximizing welfare, however expressed. Recognizing that the behavior of an individual cannot be necessarily extrapolated to the behavior of a country, many of the theoretical results from one literature are directly applicable in the other literature. The big difference is that the IEA literature emphasizes the formation of coalitions of countries to coordinate contributions to the public good; coalitions do not play a significant role in voluntary contributions to public goods by individuals.

The endogenous formation of groups or coalitions of players to coordinate the provision of public goods is in the spirit of Olson's (1971) seminal treatise on the provision of public goods. Over the past two decades, most of the work on this problem has been done in the context of the international environmental agreement (IEA). Only recently has the general public goods literature returned to addressing coalition formation (Kosfeld et al, 2009; Charness and Yang, 2014).

The literature on IEAs starts with a framework nearly identical to Eqn (1) for voluntary provision of public goods. The interesting twist added by the IEA literature (drawn from the cartel stability literature[3]) is that the noncooperative behavior is represented as a two stage game (and countries are posited to act as rational utility-maximizing agents). In the first stage (the "membership game") agents decide whether they wish to be in a coalition (an IEA). Specifically, each agent announces "in" or "out;" the first stage game generates a coalition as the Nash equilibrium in these announcements. In the second stage (the "emissions game"), the coalition acts as one and emissions choices of the coalition and fringe emerge as a Nash equilibrium in emissions conditional on the coalition formed in the first stage.

Barrett (1994) provides the first analysis of this problem in the literature (though he assumes a leader-follower structure rather than a two stage Nash equilibrium). Unfortunately,

[3] See d'Aspremont et al (1983) and Donsimoni et al (1986).

he is unable to come up with many analytic results without simplifying;[4] he uses simulations to suggest that welfare gains from an IEA (relative to the noncooperative outcome) are modest. An IEA may have many members (relative to N), but in such cases , welfare gains are slight compared to the noncooperative equilibrium; conversely, when cooperation would increase net benefits significantly, the equilibrium size of an IEA is small. In other words, generally (but not always) there is an inverse relationship between the equilibrium number of coalition members and the gains from cooperation (ie, the welfare difference between a coalitional outcome and a noncooperative outcome).[5] The intuition behind this is straightforward. An equilibrium coalition is held together by the fact that if any one player defects, the cooperative strategy of the coalition will fall apart. Thus the equilibrium coalition is the smallest one possible for which contributing to the public good (abatement) is collectively rational for the coalition. When there is more to be gained from cooperation, then the minimal viable abating coalition is smaller; when there is less to be gained, it takes more coalition members for abating to be collectively rational.

One simplification of the model (Barrett, 1999; Ulph, 2004) is for payoffs to be linear with identical preferences and identical endowments and a common MPCR. In this case, it is easy to show that the equilibrium consists of fringe agents free-riding ($g_i = 0$) and coalition members undertaking some pollution reduction (abatement), providing the coalition is large enough. Let n* be the smallest size of a coalition which chooses to abate. A basic result is that n* = 1/MPCR and all stable contributing coalitions are of this size. One can also show that the benefits from cooperation increase in MPCR whereas the size of a coalition decreases in MPCR, following the same logic as in the previous paragraph. The assumption of identical endowments is almost universal in this literature, despite the fact that the size or wealth of a

[4] Subsequent authors have refined this model, solving it model analytically. See Finus (2001), Diamantoudi and Sartzetakis (2006), and Rubio and Ulph (2006).

[5] This inverse relationship between MPCR and the equilibrium size of an IEA holds generally in Barrett (1994), as articulated in his Prop. 1. He does offer specific functional forms where it does not necessarily hold. For instance, he shows that with constant marginal damage and quadratic costs that the maximum size of an IEA is 3 members. In this special case the relationship between MPCR and IEA size will not of course hold. See also Finus (2001).

country seems, in the real world, to be an important factor in driving participation in IEAs. Few in the IEA literature have explicitly treated social preferences. An exception is Lange (2006), who explores the significance of equity in reaching agreement in an IEA. In fact, in one of the few empirical papers on this issue, Lange et al (2007) survey attitudes of individuals involved in climate negotiations and find a strong preference for equity. Whether this translates into countries caring about equity is another matter.

Carraro and Siniscalco (1993) underscore the significance of commitment mechanisms to hold coalitions together, such as punishments for defection. The problem with punishment is that it is often not individually rational for countries to penalize countries which defect. However, several authors, starting with Barrett (1999), examine credible punishments that can hold a coalition together, in the context of an infinitely repeated game. The framework is renegotiation-proof equilibria (Fudenberg and Maskin, 1986) whereby punishments are built into an agreement. Provided the discount rate is sufficiently small (agents are sufficiently patient), full cooperation can be sustained (Asheim and Holtzmark, 2009). Heitzig et al (2011) apply this to climate change mitigation.

One of the first recent papers to explore coalitions for providing public goods outside the context of international environmental agreements is Kosfeld et al (2009). In that paper, the authors suggest a stage game structure very similar to the standard static IEA problem, though with one additional stage. The first stage is a membership game, the second stage is an implementation stage and the third stage is a contribution stage. The membership and contribution stages are identical to the IEA problem. The participation stage involves members of the coalition deciding whether to "implement" the coalition. An implemented coalition involves payment of a fixed fee and punishment for not contributing enough (applied to coalition members). Their theoretical results for standard preferences are a repetition of IEA results. An interesting extension is their introduction of fairness, using Fehr-Schmidt social preferences. They show that for a subset of Fehr-Schmidt preferences, the grand coalition is an organizing equilibrium.

Researchers have only recently begun to use experiments to validate theory on the formation of coalitions for public goods provision (Kosfeld et al, 2009; Burger and Kolstad, 2009; Dannenberg et al, 2014). Results are ambiguous though consistent with the private provision literature – experimental evidence suggests more cooperation and less free-riding.

III. THEORY: THE BASIC MODEL

The classic model of international environmental agreements is the static one-shot two stage game first proposed by Barrett (1994), drawing on the earlier cartel literature (Donsimoni et al, 1986). There are other models, some of which were discussed in the review in the previous section. In the development below, it is the classic static model which we extend to involve social preferences on the part of countries (self-interest as well as prosocial objectives).

It should be emphasized that much of the literature we draw on for characterizing preferences is concerned with the behavior of individuals, not countries. At some risk, we transfer these individual-level results to the case of a country with an objective (analogous to an individual utility function). The notion that a country can be viewed as having an objective function is not embraced by some economists nor by some political scientists.

A. Basic Conditions.

Let there be i=1,…,N countries, each with the potential to emit w_i, choosing a level of abatement (a public good), g_i, or equivalently, a level of emissions, $x_i=w_i-g_i$. Welfare is positively affected by the direct benefits of emitting, x_i (including lower production costs) and positively affected by the aggregate level of abatement, G (more abatement means less environmental damage). [6] Building on the terminology of Andreoni (1990), countries are *impurely altruistic;* i.e., they have a national welfare function (u_i) that is additively separable

[6] The notation here is consistent with an alternative model, in the spirit of Bergstrom et al (1986) in which there are a group of N agents, x_i is private goods, g_i is private contributions to a public good and w_i is wealth, for agent i.

into an *egoistic* component (π_i) and a *prosocial* or *altruistic* component (α_i). National welfare can be viewed as a convex combination of these two components:

$$u_i(x_i,G) = \lambda_i \ \pi_i(x_i,G) + (1-\lambda_i)\,\alpha_i(\boldsymbol{\pi}) \qquad (2)$$

where $\lambda_i \varepsilon [0,1]$ is the parameter reflecting the extent to which the country is selfish vs. altruistic. Note that in Eqn (2), egoistic (self-centered) welfare is given by the function π, which is the welfare from the payoff from consumption ignoring the well-being of others. The altruistic (prosocial or other-regarding) component is given by the function α, which depends on the vector of egoistic payoffs for the other countries. This is a somewhat paternalistic representation of altruism in that agents care only about the egoistic payoffs of others, not the overall welfare of others. Additive separability is a modest restriction, though consistent with the literature, as can be seen below.

Rather than postulate a general egoistic welfare function which depends on individual and aggregate contributions (as in Bergstrom et al, 1986), we linearize the egoistic payoffs. This is consistent with a much of the literature, particularly in the experimental realm. It also allows us to consider heterogeneous preferences, often assumed away in more general models. In the absence of the altruistic component, this is a standard homogeneous preferences linear public goods game. Specifically, the egoistic component (which we will also refer to as the monetary payoff) is given by

$$\pi_i = x_i + aG \quad \text{where } x_i + g_i = w_i,\ G = \Sigma_i\, g_i \qquad (3a)$$

$$= w_i - g_i + aG \quad \text{where } G = \Sigma_i\, g_i \text{ and } 0 \le g_i \le w_i \qquad (3b)$$

Here w_i is maximum possible emissions for country i, g_i is the level of abatement for country i and G is the aggregate abatement over all countries. This is equivalent to a linear payoff from emissions x_i and the public good G, with country i making a contribution to a public good, g_i, subject to a feasibility constraint $x_i+g_i=w_i$. The parameter *a* is the marginal per country return

(MPCR), indicating how much of an investment in abatement is returned privately.[7] To keep the problem interesting, we restrict the MPCR, *a*, to be in the open interval (1/(N-1),1). Thus we are excluding a=1, for which the country would be indifferent between unilaterally contributed or not. We also exclude very small *a*, for which even coordination may not be enough for abatement. Clearly *a* could vary from one country to another, though that would complicate our analysis. We will assume the size of the country, in terms of potential emissions, may be different from one country to another in our group.

Although a linear payoff is common in the empirical and experimental literature, this is equivalent to assuming gains from emitting are a perfect substitute with gains from reduced damage from abating. And this inevitably leads to knife-edge outcomes wherein the agent either abates completely or not at all. Nonlinear payoffs would make for more subtle behavior but it would be more difficult to provide an analytic representation of the Nash equilibrium, particularly when heterogeneous other-regarding preferences are treated.

There are three major representations of the altruistic component of welfare in Eqn. 2 in the literature: Fehr and Schmidt (1999) [FS], Bolton and Ockenfels (2000) [BO] and Charness and Rabin (2002) [CR]. Andreoni (1990) in the empirical implementation of his model assumes altruism is manifest by introducing g_i into utility – the agent receives a warm-glow from giving. The notion of warm-glow would appear to be closely related in particular to CR, though the mechanism whereby giving to the public good generates utility is more ambiguous and undefined in Andreoni (1990).

As articulated by Cooper and Kagel (2013), both the BO and FS models define the altruistic component of welfare in terms of how an individual's payoff compares to the payoff of others. FS stipulate that $\alpha_i(\boldsymbol{\pi})$ depends on relative payoffs (for agent i) defined as:

$$\alpha_i(\boldsymbol{\pi}) = - \gamma_i/(N-1) \Sigma_{j\neq i}\max(\pi_j - \pi_i,0) - \beta_i/(N-1) \Sigma_{j\neq i}\max(\pi_i - \pi_j,0) \qquad (4a)$$

[7] This is of course more commonly termed the marginal per capita return in the context of a public goods game with individuals rather than countries. It is the same concept.

where the γ_i and β_i are in the interval [0,1) and reflect aversion to personally disadvantageous (countries doing better than i) and advantageous (countries doing worse than i) inequality, respectively (it is assumed that $\gamma_i \leq \beta_i$). The authors specifically state that even though agents may be homogenous in terms of the payoff function, some may have different attitudes towards inequality than others. Different mixes of "selfish" and "fair minded" people can result in very different levels of cooperation. We refer to Eqn. (2-4) as F-S social preferences.

BO take a simpler approach, though not restricted to a linear combination of egoistic and altruistic payoffs (as in Eqn. 2). In the linear version of their model, let

$$\alpha_i(\pi) = v_i(\pi_i / \Sigma_j \pi_j) \qquad (4b)$$

where $v_i()$ is a function of payoff share which peaks at 1/N – the point at which one's own share is equal to the average share.

Charness and Rabin (2002) suggest that efficiency is also important (see the discussion in the context of Figure 1). Although they are careful not to reject the Fehr and Schmidt representation, they suggest that it is incomplete. In fact, they suggest that utility depends on three dimensions of payoff: personal payoff, an equity term and an efficiency term. Their approach is to posit utility for agent i as a linear combination of own monetary payoff, the minimum monetary payoff over the rest of the population (a Rawlsian-like criterion reflecting a concern for equity) and total monetary payoffs over the population (reflecting concerns for social efficiency):

$$\alpha_i(\boldsymbol{\pi}) = [\, \delta_i \min_{j \neq i} \pi_j + \varepsilon_i \Sigma_j \pi_j \,]/(1 - \lambda_i) \qquad \text{where } \delta_i, \varepsilon_i \geq 0 \text{ and } \delta_i + \varepsilon_i + \lambda_i = 1 \qquad (5)$$

In Eqn. (5), δ_i reflects the relative importance to agent i of distribution/equity and ε_i reflects the importance of efficiency. We refer to Eqn. (2-3, 5) as C-R preferences, though the equity term in Eqn. (5) is slightly different from the original representation in Charness and Rabin (2002) in that the minimum excludes own payoffs:

$$u_i(g_i) = \lambda_i[\, w_i - g_i + aG)] + \delta_i[w_m - g_m + aG] + \varepsilon_i \Sigma_k [w_i - g_k + aG] \qquad (6a)$$

where $\lambda_i, \delta_i, \varepsilon_i \geq 0$, $\delta_i + \varepsilon_i + \lambda_i = 1$, $G = \Sigma_k g_k$ (6b)

and country m has the lowest payoff has the lowest payoff of all the other countries. Although modification of the minimum is in part for tractability, the fact is that a concern for equity is usually thought of as a concern for the well being of others, particularly those less well-off.

The Andreoni warm-glow model could be viewed as a variant of C-R when the warm glow arises from providing social benefits.

Because the C-R preferences appear to represent a broader perspective on social preferences (by including equity and efficiency, not just equity as in F-S), we adopt that representation here. Some readers may find the restrictions implicit in preferences given by Eqn. (2-3,5) troubling – preferences are linear and altruism is narrowly defined. However, that is how the literature, primarily experimental, has approached this problem. We adopt a form of preferences (C-R) widely used and cited in the literature.

B. Efficient and Noncooperative Outcomes

Assume C-R preferences as characterized in Eqn. (6) where individual λ, δ and ε may vary from one country to another. By assumption, a>1/N; thus the aggregate monetary (egoistic) payoff is maximized when everyone is abating as much as possible. Clearly, a Pareto optimum will occur when $g_i = w_i$ for all i and $\pi_i = u_i = aN$, for all i.

If countries are interacting non-cooperatively, we seek a Nash equilibrium. It would be helpful to apply the results of Bergstrom et al (1986) to characterize the equilibrium. However in Bergstrom et al (1986), utility of an individual agent is a function of g and G; in our case, the vector of g's enters each utility function due to the equity criterion (see Eqn. 7b below). Only if δ=0 would g_i and G be the only arguments of the utility function.

Assume country m has the lowest payoff. Thus country i≠m chooses g to maximize $u_i(g)$, defined as:

$u_i(g_i) = \lambda_i[w_i - (1-a)g_i + aG_{-i})] + \delta_i[w_m - g_m + aG] + \varepsilon_i\Sigma_k[w_i - g_k + aG]$ (7a)

$$= \lambda_i[\, w_i - (1-a)g_i + aG_{-i}] + \delta_i[w_m - g_m + a(G_{-i} + g_i)] + \varepsilon_i\,[W+(aN-1)(G_{-i}+g_i)] \qquad (7b)$$

where $W=\Sigma_k w_k$; $\lambda_i, \delta_i, \varepsilon_i \geq 0$; $\lambda_i + \delta_i + \varepsilon_i = 1$; $G_{-i} \equiv \Sigma_{j\neq i}\, g_j$. (7c)

To simplify, let $\Delta_i(g) \equiv u_i(g) - u_i(g=0)$ be the gain to i from abating, which can be written, after some simplifying, as:

$$\Delta_i(g_i) = g_i\{a[1+(N-1)\varepsilon_i]-1+\delta_i\} \equiv g_i\{a + \varepsilon_i\,[a(N-1)-1] - \lambda_i\} \qquad (8)$$

Clearly welfare in Eqn. (8) is maximized at either $g_i=0$ or $g_i=w_i$, depending on the sign of the term in braces in Eqn (8), which leads to the following proposition:

Prop. 1. Assuming the N homogeneous player public goods game with C-R social preferences (Eqn. 6), then

(1) Efficient (Pareto Optimal) outcomes involve all countries undertaking maximal abatement; and

(2) The Non-cooperative Nash equilibrium involves each agent either not abating ($g_i=0$) or fully abating ($g_i=w_i$) according to

$$g_i = 0 \text{ if } a < \bar{a}_i \qquad (9a)$$

$$g_i = w_i \text{ if } a > \bar{a}_i \qquad (9b)$$

where $\bar{a}_i = (\lambda_i+\varepsilon_i)/[1+\varepsilon_i(N-1)]$ (9c)

In the case where $\bar{a}_i = a$, then any feasible abatement level for country i is a Nash equilibrium.

The intuition behind Prop. 1 is straightforward. With standard egoistic preferences, the cutoff between abating and not is a=1. With social preferences, the cutoff is lower: $\bar{a} \leq 1$. Eqn. (9c) simply defines how social preferences reduce this cutoff.

Note in Prop. 1 that if $\lambda_i=1$ (standard preferences—all weight is on egoistic payoffs), then $\delta_i = \varepsilon_i = 0$ (by Eqn. 7c) and $\bar{a}_i=1$: the Nash equilibrium is for all countries to contribute

nothing to the public good, since by assumption $a_i < 1$. The $\bar{a}_i$ in Prop. 1 can be interpreted as the cutoff MPCR (varying from country to country) between cooperation and noncooperation. The effect of other-regarding social preferences with some concern for efficiency ($\varepsilon_i > 0$), keeping δ_i constant, is to lower $\bar{a}_i$, effectively expanding the levels of MPCR wherein cooperation takes place. This result is quite similar to Proposition 4 in Fehr and Schmidt (1999), a theorem in which the authors expand the set of MPCRs for which contributing is individually rational. Note further that in Prop. 1 above, when efficiency is of some concern, then increasing N has the effect of lowering $\bar{a}_i$. The logic is simply that from an efficiency point of view, the payoff from abating increases as the number of countries increases. When $\varepsilon_i=0$ (welfare does not depend on efficiency), N drops out of $\bar{a}_i$ and the effect of N on $\bar{a}_i$ disappears:

Corollary 1. Assuming the N homogeneous country abatement game with C-R social preferences Eqn. (6), then the cutoff level between noncooperation and cooperation for an individual country ($\bar{a}_i$) as defined in Prop 1, exhibits the following comparative statics:

a. If $\varepsilon_i > 0$, then increasing the number of countries (N) has the effect of lowering the cutoff MPCR level between cooperation and noncooperation ($\bar{a}_i$), effectively shrinking the range of values of MPCR associated with noncooperation.

b. If $\varepsilon_i = 0$ (no concern for efficiency), then changing the number of countries (N) has no effect on the cutoff MPCR level between cooperation and noncooperation ($\bar{a}_i$).

c. Decreasing λ (holding either ε or δ fixed) has the effect of reducing the cutoff MPCR.

IV. ENDOGENOUS COALITIONS FACILITATING PROVISION

We now consider a slightly more complicated institution. We allow a subset of countries to endogenously form a coalition for the express purpose of coordinating abatement—an international environmental agreement (IEA). This is in the spirit of the groups explored by Olson (1971). Countries voluntarily join the coalition and may leave the coalition.

Furthermore, any abatement provided by coalition members benefits both coalition members and non-members (thus it is not a club good in the standard sense). This leads to the obvious question: why would anyone join the coalition when the fringe enjoys all of the benefits and none of the costs of the coalition? The answer to this legitimate question lies in the nature of a Nash equilibrium. A Nash equilibrium is an allocation wherein it is not in any agent's individual interest to unilaterally change behavior. Some countries find themselves in the coalition in equilibrium (and some not), with no incentive to unilaterally defect. It is not relevant what path a country took to find itself in the coalition or in the fringe (or even, in fact, if such a path exists).

As is standard in the literature on cartels and international environmental agreements, we view the problem as a two stage game. In the first stage, countries decide whether or not to join the coalition. In the second stage, countries decide how much to abate, with the coalition acting as one – abating at a joint payoff maximum for the coalition. We solve the problem using backwards induction.

Before moving to these two stages, some notation is in order. Define the members of the coalition by C = {i|country i is a member of the coalition} and the number of members of the coalition, C, by n. Let W_C be the aggregate potential emissions of the coalition and G_C be the total abatement from the coalition members and G_F the total abatement from the fringe.

A. Abatement Stage

In the two-stage game, the second stage is the abatement choice stage, when countries in the fringe and the coalition determine how much to abate, conditional on the size and composition of the coalition. This leads to our first result regarding the actions of the fringe.

Lemma 1. With homogeneous C-R preferences and countries divided into members of the coalition and members of the fringe, it is a dominant strategy for each member (i) of the fringe to abate maximally or not at all, depending on the value of $\bar{a}_i$ relative to a:

Abate w_i if $a > \bar{a}_i$ (10a)

Abate 0 if $a < \bar{a}_i$ (10b)

where $\bar{a}_i$ is defined by Eqn 9c.

Pf: Identical to proof of Prop. 1 □

We can similarly examine the incentives of the coalition, though we need to slightly restrict the minimum payoff in the C-R preferences. As defined, the minimum is over all agents in the economy. However, the coalition will be aggregating utility over members of the coalition. Thus it makes more sense (and is more tractable mathematically) to view equity within the coalition with respect to the minimum payoff country *outside of the coalition*. It is unlikely that this is a significant restriction:

Lemma 2. Assume C-R preferences and countries divided into members of the coalition and members of the fringe. Furthermore, in the C-R preferences, assume equity concerns of coalition members are with respect to the minimum payoff of countries outside the coalition. Then, conditional on the size of the coalition being n, with coalition members indexed by C, it is a dominant strategy for member i of the coalition to either abate at the level w_i or not at all, according to:

$$g_i = 0 \text{ if } a < \bar{a}_i(C) \quad (11a)$$

$$g_i = w_i \text{ if } a > \bar{a}_i(C) \quad (11b)$$

where $\bar{a}_i(C) \equiv [\lambda_i + \Sigma_{k\epsilon C}\, \varepsilon_i] / \Sigma_{k\epsilon C}\, [1+\varepsilon_k(N-1)]$ (11c)

Pf: The aggregate payoff for the members of the coalition, $k\epsilon C$, when individual abatement is g_k:

$$\Pi_C(\mathbf{g}) = \Sigma_{k\epsilon C}\, \{\lambda_k(w_k - g_k + a(G_C+G_F)) + \delta_k(w_m + a(G_C+G_F)) + \varepsilon_k \Sigma_i (w_i - g_i + a(G_C+G_F))\}$$

$$= \Sigma_{k\epsilon C}\, \{\lambda_k(w_k - g_k + a(G_C+G_F)) + (w_m + a(G_C+G_F))\, \Sigma_{k\epsilon C}\, \delta_k + \Sigma_{k\epsilon C}\, \varepsilon_i\, [W + a(N-1)\,(G_C+G_F)] \quad (12)$$

where G_F and G_C are the aggregate levels of abatement in the fringe and coalition, respectively. Thus the payoff for the coalition undertaking no abatement is

$$\Pi_C(\mathbf{0}) = \Sigma_{k\epsilon C}\{\lambda_k(w_k+aG_F) + (w_m+aG_F)\Sigma_{k\epsilon C}\delta_k + \Sigma_{k\epsilon C}\varepsilon_i[W+a(N-1)G_F] \quad (13)$$

which implies that the difference in payoff between contributing and not, $\Delta(\mathbf{g}) \equiv \Pi_C(\mathbf{g}) - \Pi_C(\mathbf{0})$ is

$$\Delta(\mathbf{g}) = -\Sigma_{k\epsilon C}\lambda_k g_k + G_C\Sigma_{k\epsilon C}\{a + \varepsilon_i[a(N-1)-1]\} . \quad (14)$$

Clearly the rhs of Eqn. (14) is zero when $G_C=0$ (at $\mathbf{g}=\mathbf{0}$). Payoffs will increase for abatement from any k for which $d\Delta/dg_k > 0$:

$$d\Delta/dg_k = -\lambda_k + \Sigma_{k\epsilon C}\{a + \varepsilon_i[a(N-1)-1]\} > 0 \quad (15a)$$

$$\Leftrightarrow a > [\lambda_k + \Sigma_{k\epsilon C}\varepsilon_i]/\Sigma_{k\epsilon C}[1+\varepsilon_i(N-1)] = \bar{a}_i(C) \quad (15b)$$

Note that the term in braces in Eqn. (15a) is always positive (by assumption). Thus for any k for which Eqn. (15) holds, net payoffs will be increased by abating, implying that $g_k=w_k$. For any k for which Eqn. (15) fails to hold (leaving aside cases of equality), abating will only reduce welfare; thus $g_k=0$. This completes the proof. □

Note the similarity between Lemma 2 and the noncooperative decision to abate in Prop. 1. There are two primary differences between the two results. One is that the size of the denominator is increased by virtue of the size of the coalition (even if ε is zero). This is a standard result in the standard theory of international agreements. However, in addition, the concern for efficiency embodied in ε, the effect is magnified by the aggregation of ε's over the coalition. The intuition will become clearer in the next section when we consider a simplification involving identical preferences.

The following proposition follows directly from these results:

Prop. 2. Assume C-R preferences and countries divided into members of the coalition and members of the fringe. Furthermore, in the C-R preferences, assume equity concerns of coalition members are with respect to the minimum payoff of countries outside the coalition.

Then, conditional on the size of the coalition being n, with coalition members indexed by C, it is a dominant strategy for the coalition to abate G_C:

$$G_C = \Sigma_{k\epsilon R}\, w_k, \text{ where } R = \{k\epsilon C \mid \bar{a}_i(C) < a\}, \qquad (16)$$

where $\bar{a}_i(C)$ is defined in Eqn. (11c).

Pf: The proof is a direct application of Lemma 2. □

B. Membership Stage.

Nash equilibrium in the membership game is easy to define but hard to connect to fundamental characteristics of the game for the case of general heterogeneous preferences. In particular, a coalition C is a Nash equilibrium of the membership game if (a) the payoff attained by any member of the coalition, d, inside the coalition is as great as that country can obtain outside the coalition when the coalition is reduced to C-{d}; and (b) no member of the fringe, f, outside the coalition can do better by joining the coalition, making it C U {f}. In other words, it is individually rational for each member of the fringe to stay in the fringe and for each member of the coalition to stay in the coalition. This is simply the definition of a Nash equilibrium. For simplicity, we focus on the case where social preferences are the same among countries (λ, δ and ε), though endowments may differ. We consider this case in the next section.

V. IDENTICAL PREFERENCES, HETEROGENEOUS SIZE

In the previous sections, we developed theory for the general case of heterogeneous countries with heterogeneous levels of potential emissions. Although it is possible to obtain conditions charcterizing an equilibrium in the coalition game for this general problem, clarity is not well served. It is common (eg, see Bergstrom et al, 1986) to assume homogeneous preferences and let heterogeneity be manifest through different sizes (potential pollution). We consider that special case here. Thus we assume all countries share the same λ, δ and ε, though have different levels of potential pollution, w_i.

The first result concerns the abatement stage of the two stage game. Focusing on the case where the noncooperative equilibrium involves no abatement, there is a clear result on the size of a coalition necessary to support abatement as an optimal strategy for the coalition:

Prop. 3. Assume homogeneous C-R preferences such that a non-cooperative equilibrium yields no abatement. Define ñ as:

$$\tilde{n} = \text{ceil}[1/\bar{e}], \tag{17a}$$

$$\text{where } \bar{e} \equiv \{a + \varepsilon[a(N-1)-1]\}/\lambda \tag{17b}$$

and the function ceil(x) is smallest integer greater than or equal to x. Then for a two-stage public goods game with a coalition C, of size n and collective size W_C, it is collectively rational for the coalition (i.e., in the coalition's aggregate best interest) to abate $G_C = W_C$ if $n \geq \tilde{n}$ and $G_C=0$ otherwise.

Pf: From Eqn. (11c), in Prop. 2,

$$\bar{a}_i(C) \equiv [\lambda_i + \Sigma_{k\epsilon C}\, \varepsilon_i]/\, \Sigma_{k\epsilon C}\, [1+\varepsilon_k(N-1)] \tag{18a}$$

which, for homogeneous preferences, equivalent to

$$\bar{a} \equiv [\lambda + n\varepsilon]/\, n[1+\varepsilon(N-1)] \quad . \tag{18b}$$

From that Proposition, it is collectively rational for the coalition to fully abate if $a < \bar{a}$ (and similarly to not abate at all if $a > \bar{a}$). The condition that $a < \bar{a}$ can be combined with Eqn. (18b) to yield a condition for it being in the best interest of the coalition to fully abate:

$$n\bar{e} < \lambda \tag{18c}$$

which proves the proposition. □

We now turn our attention to the incentives of individual members of the coalition to voluntarily abate. We are interested in the conditions for abatement being *individually rational* for members of the coalition. This is equivalent to internal stability. If it is in the individual

interests of a member of the coalition to not abate, then that is equivalent to joining the fringe. It turns out that there all equilibrium coalitions are of the same size – too large and members defect, too small and abatement is not collectively rational (nor individually rational). We first prove a result on individual rationality.

Lemma 3: Assume homogeneous C-R preferences such that a non-cooperative equilibrium yields no abatement and with ñ defined as in Eqn. (17). Then for a two-stage public goods game with a coalition C, of size n, with the coalition acting collectively rationally, it is individually rational for coalition members to abate fully if n is such that $\tilde{n} \leq n < \tilde{n} + 1$; otherwise it is individually rational for each coalition member to pollute and not abate.

Pf: For $n < \tilde{n}$, we already know from Prop. 3 that the coalition will choose to pollute and by assumption, individuals will not want to unilaterally abate. We thus focus on the situation where the coalition finds abating collectively rational ($n \geq \tilde{n}$). In particular, we consider two cases, one in which $n \geq \tilde{n} + 1$ (in which case we show that polluting is individually rational) and one in which $\tilde{n} \leq n < \tilde{n} + 1$, where abating is individually rational, conditional on the other members of the coalition pursuing the collectively rational strategy.

From Eqn. (7), the payoff for country i in the coalition is given by

$$u_i(g_i) = \lambda[w_i - (1-a)g_i + aG_{-i})] + \delta_i[w_m - g_m + aG_C] + \varepsilon\Sigma_k[w_i - g_k + aG_{-i}]$$

$$= \lambda[w_i - (1-a)g_i + aG_{-i}] + \delta[w_m - g_m + a(G_{-i} + g_i)] + \varepsilon[W + (aN-1)(G_{-i} + g_i)] \qquad (19a)$$

For the case of $n \geq \tilde{n} + 1$, the coalition will continue to want to abate, with or without a defector. Thus aggregate abatement will be G_C for the entire coalition and $G_C - w_i$, should i defect and provide no abatement. This means that the gains for country i from defecting -- polluting rather than fully abating -- are given by

$$\Delta_i = \lambda[w_i + a(G_C - w_i)] + \delta[w_m - g_m + a(G_C - w_i)] + \varepsilon[W + (aN-1)(G_C - w_i)]$$

$$- \lambda aG_C - \delta[w_m - g_m + aG_C] - \varepsilon[W + (aN-1)(G_C)]$$

$$= \lambda(1-a)\, w_i - \delta a w_i - \varepsilon(aN-1)w_i$$

$$= \{\lambda + \varepsilon - a[1 + \varepsilon(N-1)]\}w_i \qquad (20)$$

By assumption the noncooperative equilibrium involves no abatement; thus from Prop 1, we know Δ_i in Eqn (20) is positive and thus that it is individually rational to defect from the coalition decision.

For the case of $\tilde{n} \leq n < \tilde{n} + 1$, we know that should one member defect, the coalition will no longer choose to abate since it will be too small (by Prop. 3). Thus the gain from defecting is given by

$$\Delta_i = \lambda w_i + \delta w_m + \varepsilon W$$

$$- \lambda a G_C - \delta[w_m + a G_C] - \varepsilon[W + (aN-1)G_C]$$

$$= \lambda(w_i - a G_C) - \delta\, a\, G_C - \varepsilon\,(aN-1)G_C$$

$$= \lambda\, w_i - G_C\{a[1+\varepsilon(N-1)] - \varepsilon\}$$

$$< G_C\{\lambda + \varepsilon + a[1+\varepsilon(N-1)]\} \qquad (21)$$

where the inequality in Eqn. (21) is from the fact that $w_i < G_C$. Since by assumption $a > \bar{a}$, then by Prop. 1, the right-hand-side of Eqn. (21) is negative which implies that it is not individually rational to defect from the coalition. □

It is straightforward to put these results together into a proposition regarding the solution of the two-stage game.

Prop. 4. Assume homogeneous C-R preferences such that a non-cooperative equilibrium yields no abatement. Consider the two stage game involving a first stage membership game and a second stage abatement game with an equilibrium involving a Nash equilibrium in both stages. Allow coalitions to form in the membership stage and define $\tilde{n}$ as in Eqn. (17). All coalitions of

size ñ are both internally and externally stable (ie, they are Nash equilibria for the two stage game); all coalitions of different size are internally unstable.

Pf: External stability is easy, since by assumption it is a dominant strategy f 0)or individual countries to pollute. Internal stability follows from Lemma 3. □

The standard linear model for coalition formation involves identical endowments and monetary payoffs. The main result from that literature is that all stable coalitions have size ceil(a). One obvious question is how moving from pure self-interested standard preferences ($\lambda=0$) to other regarding preferences changes the cutoff size of a coalition in Prop. 3. This is easily answered with comparative statics:

Prop. 5. Assuming homogeneous C-R preferences as in Prop. 4, then

(1) $d\tilde{n}/d\lambda \geq 0$, holding either δ or ε constant;

(2) $d\tilde{n}/d\delta$ is ≥ 0 holding λ constant and ≤ 0, holding ε constant; and

(3) $d\tilde{n}/d\varepsilon$ is ≥ 0 holding δ constant and ≤ 0, holding λ constant.

Pf: Combine Eqn (17a) and Eqn (17b) and then totally differentiate, ignoring the ceil function. Further, totally differentiate the identity $\lambda + \delta + \varepsilon = 1$. Solve the two equations for $d\tilde{n}/d\lambda$, $d\tilde{n}/d\varepsilon$, or $d\tilde{n}/d\delta$, which proves the lemma. □

This Proposition indicates that introducing C-R social preferences (lowering λ) tends to lower the size of stable coalitions. Introducing size heterogeneity among countries does not change the size of coalitions but does influence the overall level of abatement (Prop. 4). This result on the size of coalitions does contradict the results of Kosfeld et al (2009) who show that with Fehr-Schmidt preferences (in which players dislike payoff inequality), larger coalitions may be equilibria (even the grand coalition), depending on parameter values.

It is common to interpret the result on the size of stable coalitions as a "dismal" result in the theory of IEA's – dismal in the sense that many MPCRs lead to very small coalitions and the

larger the MPCR, the smaller the coalition (Barrett, 1994). We would argue that the more appropriate interpretation of the size of stable coalitions is the size of the *smallest effective coalition*. In reality, additional tools will be used to keep a coalition together – incentives, punishments and interconnecting coalitions, to name a few. However, it is difficult to overcome the situation wherein a coalition is so small that it is not in the self-interest of the coalition to contribute to the public good. Thus the fact that social preferences tend to shrink the size of an effective coalition to provide abatement is good news. Viewing preferences as social can expand the set of viable coalitions.

It is appropriate to compare these results with those of the simple public goods game without coalition formation but with heterogeneous agents, some of whom wish to unilaterally abate and some of whom do not. In this case, there will be a subgroup of contributors and a subgroup of noncontributors. The contributors may appear to be a coalition of agents with the same agenda; however the theory really only suggests that individuals in the group will individually contribute. For example, many argue that the Montreal Protocol for protecting stratospheric ozone is a good example of a coalition of nations to provide abatement. However, Murdoch and Sandler (1997) have suggested that the Montreal Protocol is largely an association of countries which find reducing emissions individually rational.

Note that any coalition of size ñ is an equilibrium coalition. Of all the possible coalitions of size ñ, the one with the highest aggregate size, W_C, will yield the maximal amount of abatement, G_C. Thus from an efficiency point of view there is one coalition which is best (assuming there are not too distinct coalitions of the same size). Coalitions of large countries are better, from an efficiency perspective, than coalitions of smaller countries.

VI CONCLUSIONS

In this paper we revisit the important question of self-enforcing international environmental agreements. In particular, we are interested in two issues. One is the role of

social preferences—where self-interest is not the only motivator. How do social preferences change the received wisdom on abatement in the absence of an IEA? The second issue is the role of voluntary coalitions in coordinating the provision of abatement. Little is known of how social preferences modify what we know about international environmental agreements (viewed as abstract economic coordination entities).

We adopt the specification of social preferences due to Charness and Rabin (2002), primarily because it contains three important ingredients that characterize many discussions of social preferences: private gain, equity and social efficiency. Using a linear public goods model with a fixed MPCR but an arbitrary distribution of size (proxied as potential emissions), we find that a major consequence of using social preferences is to lower the threshold for abatement being individually rational. This is in contrast to theory with standard preferences where for any MPCR less than one, free-riding is individually rational. Thus with heterogeneous preferences, more countries may find it individually rational to abate (provide the public good) with social preferences as compared to standard self-interested preferences.

In extending the analysis to voluntary coalitions, we show that social preferences tend to reduce the size of an equilibrium coalition. Although this would appear to be a discouraging result, another interpretation is that social preferences expand the set of coalitions for which it is collectively rational for the coalition to provide the abatement public good. By implication, the smallest collectively rational contributing coalition is smaller than it would be with standard preferences: size matters, but in an unusual way.

Related to this result is the fact that conditions that determine the equilibrium size of a stable coalition are independent of the size of countries. Yet the aggregate abatement provided by a country or coalition is influenced by size. This suggests that as a refinement of the multiple stable coalitions that are possible, coalitions of the largest countries will generate the maximum amount of abatement, and thus the greatest amount of social surplus.

REFERENCES

Andreoni, J., and J. Miller. “Giving according to GARP: an experimental test of the consistency of preferences for altruism.” Econometrica, 70:737–53 (2002).

Andreoni, James, “Privately Provided Public Goods in a Large Economy: The Limits of Altruism,” *J. Pub. Econ.*, **35**:57-73 (1988).

Andreoni, James, “Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence,” *J. Pol. Economy*, **97**:1447-58 (1989).

Andreoni, James, “Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving,” *Economic J*., **100**:464-77 (1990).

Asheim, GB and B. Holtzmark, “Renegotiation-Proof Climate Agreements with Full Participation: Conditions for Pareto-Efficiency,” Env. and Res. Econ., 43:519-33 (2009).

Barrett, Scott, “Self-Enforcing International Environmental Agreements,” *Oxford Economic Papers*. **46**:878-94 (1994).

Barrett, Scott, “A Theory of Full International Cooperation,” *J. Theoretical Politics*, **11**:519-41 (1999).

Becker, Gary, “A Theory of Social Interactions,” *J. Pol. Econ*., **82**:1063-93 (1974).

Bergstrom, Ted, Larry Blume and Hal Varian, “On the Private Provision of Public Goods,” *J. Pub. Econ*., **29**:25-49 (1986).

Bliss, C and B. Nalebuff, “Dragon Slaying and Ballroom Dancing – The Private Supply of a Public Good,” *J. Pub. Econ.*, **25**:1-12 (1984).

Bolton, Gary E., and Axel Ockenfels, “ERC: A Theory of Equity, Reciprocity and Competition,” Amer. Econ. Rev., 90:166-93 (2000).

Burger, Nicholas and Charles D. Kolstad, “Voluntary Public Goods Provision, Coalition Formation and Uncertainty,” NBER Working Paper 15543, Cambridge, Mass. (Nov. 2009).

Carraro, C. and D. Siniscalco, “Strategies for the International Protection of the Environment.” *J Public Econ.*, **52**:309-28 (1993).

Charness, Gary and Mathew Rabin, “Understanding Social Preferences with Simple Tests,” *Quart. J. Econ*., **117**:817-69 (2002).

Charness, G., L. Rigotti, and A. Rustichini, “Cooperation rates as a function of payoffs for mutual cooperation," UC Santa Barbara Working Paper (2008).

Charness, Gary and Chun-Lei Yang, “Starting Small towards Voluntary Formation of Efficient Large Groups in Public Goods Provision,” *J. Econ. Beh. Org,* 102:119-32 (2014).

Chaudhuri, Ananish, “Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature,” *Exp. Econ*. 14:47-83 (2011).

Cooper, D.J. and J. H. Kagel, “Other-Regarding Preferences: A Selective Survey of Experimental Results,” Handbook of Experimental Economics (forthcoming, 2013).

d’Aspremont, C., A. Jacquemin, J. Jaskold-Gabszewicz and J. Weymark, “On the Stability of Collusive Price Leadership,” *Canadian J. Economics*, **16**:17-25 (1983).

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

Diamantoudi, E. and E.S. Sartzetakis, “Stable International Environmental Agreements: An Analytical Approach,” *J. Public Econ. Theory*, **8**:247-63 (2006).

Donsimoni, M.-P., N.S. Economides, and H.M. Polemarcharkis, “Stable Cartels,” *International Economic Rev*., **27**:317-27 (1986)

Farrell, Joseph and Eric Maskin, "Renegotiation in repeated games," *Games and Economic Behavior*, 1:327-360 (1989)

Fehr, Ernst and Klaus M. Schmidt, “A Theory of Fairness, Competition and Cooperation,” *Q. J. Econ*., **114**:817-68 (1999).

Finus, Michael, *Game Theory and International Environmental Cooperation* (Edward Elgar, Cheltenham, 2001).

Fischbacher, Urs and Simon Gächter, “Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments,” *Amer. Econ. Rev*., **100**:541-56 (2010).

Fischbacher, Urs, Simon Gächter, and Ernst Fehr, “Are People Conditionally Cooperative? Evidence from a Public Goods Experiment,” *Econ. Letters*, **71**:397-404 (2001).

Heitzig, J., K. Lessmann and Y. Zou, “Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games,” Proc. Natl Acad Sci, 108:15739-44 (2011).

Isaac, R.M. and J. M. Walker, “Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism.” *Quart. J. Econ.*. **103**:179-99 (1988).

Isaac, R.M., J.M. Walker and A.W. Williams, "Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups." *J Pub. Econ.,*. **54:**1-36 (1994).

Kahneman, Daniel, Jack L. Knetsch and Richard Thaler, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *Amer. Econ. Rev*., **76**:728-41 (1986).

Kim, O. and J.M. Walker, "The Free Rider Problem: Experimental Evidence." *Public Choice*. **43**:3-24 (1984).

Kosfeld, Michael, Akira Okada, and Arno Riedl, "Institution Formation in Public Goods Games," *Amer. Econ. Rev.,* **99**:1335-55 (2009).

Lange, Andreas and Carsten Vogt, "Cooperation in international environmental negotiations due to a preference for equity," J. Public Econ., 87:2049-2067 (2003).

Lange, Andreas, "The Impact of Equity Preferences on the Stability of International Environmental Agreements," *Env. Res. Econ*., **34**:247-67 (2006).

Ledyard, John O., "Public Goods: Some Experimental Results," Ch. 2 in J. Kagel and A. Roth (Eds), *Handbook of Experimental Economics* (Princeton University Press, Princeton, NJ, 1995).

Murdoch, J.C. and T. Sandler, "The Voluntary Provision of a Pure Public Good: The Case of Reduced CFC Emissions and the Montreal Protocol," *J. Pub. Econ*., **63**:331-49 (1997).

Olson, Mancur, *The Logic of Collective Action*, 2nd Ed. (Harvard University Press, Cambridge, Mass., 1971).

Rubio, S. J. and A. Ulph, "Self-Enforcing International Environmental Agreements Revisited," *Oxford Econ. Papers*, **58**:233-63 (2006).

Smith, Vernon L., "Experiments with a Decentralized Mechanism for Public Goods Decisions," *Amer. Econ. Rev*., **70**:584-99 (1980).

Sobel, Joel, “Independent Preferences and Reciprocity,” *J. Econ. Lit*., **43**:392-436 (2005).

Ulph A., “Stable international environmental agreements with a stock pollutant, uncertainty and learning,” *J. Risk and Uncertainty*, **29**:53-73 (2004).