International Environmental Agreements with Endogenous or Exogenous Risk*

Fuhai Hong† Larry Karp‡

August 27, 2013

Abstract

We examine the effect of endogenous and exogenous risk on the equilibrium (expected) membership of an International Environmental Agreement when countries are risk averse. Endogenous risk arises when countries use mixed rather than pure strategies at the participation game, and exogenous risk arises from the inherent uncertainty about the costs and benefits of increased abatement. Under endogenous risk, an increase in risk aversion increases expected participation. Under exogenous risk and pure strategies, increased risk aversion weakly decreases equilibrium participation if and only if the variance of income decreases with abatement.

Keywords: International Environmental Agreement, climate agreement, endogenous risk, exogenous risk, risk aversion, mixed strategy.

JEL classification numbers: D8, H4, Q54

*We thank Vincent Boucher and Yann Bramoulle for comments on an earlier version of this paper.

†Division of Economics, Nanyang Technological University. Email: fhhong@ntu.edu.sg.

‡Department of Agricultural and Resource Economics, University of California, Berkeley, and the Ragnar Frisch Center for Economic Research. Email: karp@berkeley.edu.

1 Introduction

There are at least two types of risk associated with climate negotiations. We do not know whether nations will sign a treaty that leads to meaningful action, and in the event that such a treaty does emerge, we do not know the net benefit of those actions. The first type of risk depends on the interaction amongst nations and is therefore endogenous to negotiations. The second type is exogenous, because it is due to imperfect knowledge and inherent risk, e.g. about the costs and benefits of reducing greenhouse gas (GHG) emissions. For risk averse countries, we consider the effect of both types of risk on equilibrium participation in an International Environmental Agreement (IEA). Exogenous and endogenous risk have different implications for the (expected) level of equilibrium participation in an IEA.

Our paper is the first to examine the effect of risk aversion on equilibrium IEA participation under endogenous risk. We also extend and clarify Boucher and Bramoulle (2010), who study the effect of risk aversion with exogenous risk. We examine the two types of risk in isolation, because combining them into a single model adds complexity without producing additional insights.

Benedick (2009, page xv), a prominent US negotiator for many IEAs, notes academics' tendency to view the negotiating process as mechanistic, yielding a deterministic outcome. He emphasizes the contingency of the process, and the possibility of surprises. To capture this endogenous uncertainty, Hong and Karp (2012) assume that countries use mixed rather than pure strategies when deciding whether to participate in an IEA. For risk neutral countries, they show that expected participation under mixed strategies is lower than participation under pure strategies, except for a narrow range of parameter space where unilateral abatement is "almost" a dominant strategy. Mixed strategies create endogenous risk.[1] We extend the results in Hong and Karp (2012), showing that risk aversion increases the equilibrium mixed strategy

[1]Several papers study different comparative static relations involving mixed strategies and risk aversion. Engelmann (2003) and Engelmann and Steiner (2007) show that increased risk aversion in 2×2 games can increase equilibrium payoffs. Collins and Sherstyuk (2000) find that risk aversion leads to more dispersion in the mixed strategy equilibrium to a game in which firms select their location. Chuah et al (2011) examine risk aversion's effect on escalation bargaining.

participation probability; for sufficiently risk averse countries, equilibrium expected membership is greater than the deterministic pure strategy level.

Boucher and Bramoulle (2010) consider the effect, on equilibrium IEA participation, of risk aversion in the presence of exogenous risk.[2] Their abstract states

> When countries directly contribute to a public good, uncertainty tends to lower signatories' efforts but may increase participation... In contrast, when countries try to reduce a global public bad, uncertainty tends to increase signatories' efforts and decrease participation."

The authors treat climate change as a problem of reducing a public bad, and they conclude that risk aversion tends to decrease participation. This distinction between public goods and bads is misleading: an outcome does not depend on whether we define the action as abatement (a public good) or emissions (a public bad). Although we disagree with their interpretation of their own results, their formal analysis is correct and valuable. We provide an alternative analysis of the participation game under exogenous risk. The equilibrium effect of risk aversion has – contrary to their claim – nothing to do with whether the action is a public good or a public bad, and – consistent with their results – everything to do with the manner in which exogenous uncertainty influences the effect of an action on payoff volatility.

2 Model basics

Barrett (1999) first proposed the following IEA model; see also Barrett (2003, Chapter 7), Burger and Kolstad (2009) and Kolstad (2011) (chapter 19).[3] There are N

[2] With exogenous risk, Endres and Ohl (2003) show that risk aversion may increase prospects of cooperation in transboundary pollution. Bramoulle and Treich (2009), which we discuss below, show that risk aversion reduces GHG emissions in a non-cooperative equilibrium. Neither paper considers equilibrium participation in an IEA.

[3] The Nash equilibrium to this game achieves only a small fraction of potential gains to cooperation. Karp and Simon (2013) show that this conclusion can be overturned with more general functional forms, and in that respect is fragile. We adopt this functional form because of its tractability for comparative statics exercises, without relying on its implications concerning the equilibrium level of welfare.

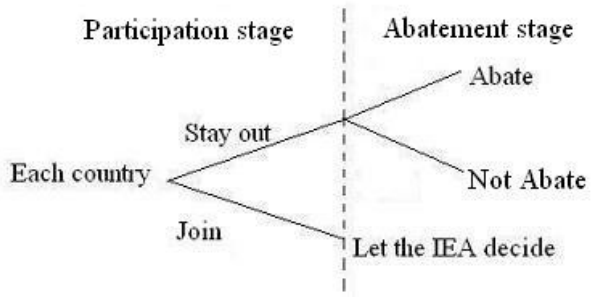

Figure 1: A canonical IEA model

identical countries, each of which has two sequential decisions: participation in an IEA and abatement. The equilibrium is subgame perfect. The cost and benefit of abatement are both linear, so in equilibrium each country abates at the level 0 or at capacity, normalized to 1. Given linearity and the absence (here) of exogenous risk, there is no additional loss in generality in assuming that the abatement decision is binary. We adopt this assumption here and in Section 3 and relax it in section 4, where we introduce exogenous risk. Abatement is a global public good. By choice of units, the benefit of each unit of abatement, to each country, equals 1. Each country's abatement cost is c, with $1 < c < N$. The first inequality means that it is a dominant strategy for a country acting alone not to abate, and the second inequality means that the world is better off when a country abates. Figure 1 shows the two stages of the game. In the first stage, each country decides whether to participate in an IEA. In the second stage, non-members choose their individually rational level of abatement, here equal to 0 because $1 < c$.

The IEA maximizes the total welfare of its members, instructing all members whether to abate. Conditional on membership k, the IEA instructs its members to abate if and only if $k - c \geq 0$. An IEA with $f(c)$ members, where the function $f(x)$ returns the smallest integer not less than x, is the "minimally successful IEA" under the assumption that the IEA optimizes conditional on membership.

Any outcome with more than $f(c)$ members is not an equilibrium in pure strategies because the "extra" members would want to leave: each member's defection

leaves unchanged other members' equilibrium action, resulting in a net benefit $c-1 > 0$ to the defector. An outcome with $f(c)$ members is an equilibrium; defection by a single member causes the IEA to instruct remaining members to not abate, causing a net loss to the defector $f(c)-c \geq 0$; a non-member loses $c-1$ by joining. Under the tie-breaking assumption that a country indifferent between joining and staying out of the IEA, decides to join, $f(c)$ is the unique pure strategy equilibrium, in addition to being the minimally successful IEA. For example, with $c = 3.2$, there are four members in this equilibrium.

3 Endogenous risk

A pure strategy equilibrium does not capture the risk of nations failing to agree on a meaningful treaty; a mixed strategy equilibrium in the participation game captures exactly this risk. The use of mixed rather than pure strategies (with no exogenous risk) gives rise to endogenous risk about participation. IEA membership is a realization of a random variable; it may or may not reach the critical level above which the IEA instructs its members to abate. We show that risk aversion increases the equilibrium expected level of participation.

Assume that countries randomize their participation decisions. Recall that the IEA instructs its members to abate if and only if there are at least $k = f(c)$ members. Therefore, a country is pivotal if and only if exactly $f(c) - 1$ other countries join. If fewer than $f(c) - 1$ other countries join, then there would still be too few members to elicit abatement even if an additional country joins. If more than $f(c) - 1$ other countries join, then membership by an additional country is (from its own standpoint) harmful: by joining it has no effect on other countries' abatement but it incurs the net cost $c - 1 > 0$. Using the abbreviation $f = f(c)$ and denoting the probability of joining as p, the probability that a country is pivotal is

$$g(p) = \frac{(N-1)!}{(f-1)!\,(N-f)!} p^{f-1} (1-p)^{N-f}.$$

The probability that at least f other countries join is

$$G(p) = \sum_{i=f}^{N-1} \frac{(N-1)!}{i!\,(N-1-i)!} p^i (1-p)^{N-1-i}.$$

Under risk neutrality, the net expected benefit of joining is the benefit of joining when the country is pivotal, $f-c$, times the probability that it is pivotal, g, minus the loss of joining when the IEA would have abated even had this country not joined, $c-1$, times the probability of that event, G. In a mixed strategy equilibrium, p must be such that a country is indifferent between joining and not joining. The equilibrium condition for p under risk neutrality is therefore

$$(f-c)\, g(p) = (c-1)\, G(p), \tag{1}$$

which is equation (1) of Hong and Karp (2012).

To examine the effect of risk aversion, define A as a country's baseline income in the absence of environmental damage or abatement, N as the environmental damage when no country abates, and let $y^m(k)$and $y^n(k)$ be, respectively, income of a member and a non-member of an IEA with k members.[4]

$$\begin{aligned} y^m(k) &= \left\{ \begin{array}{c} A-(N-k+c) \text{ if } k \geq c \\ A-N \text{ if } k<c \end{array} \right\}, \\ y^n(k) &= \left\{ \begin{array}{c} A-(N-k) \text{ if } k \geq c \\ A-N \text{ if } k<c \end{array} \right\}. \end{aligned}$$

Let $U(y)$ be a concave function that is strictly concave in a subset of the interval $[A-N+f-c, A-1]$. With preferences $U(y)$, countries are risk averse, and strictly risk averse over some interval. In the participation game, the utility gain of joining when the country is pivotal is $U(A-N+f-c) - U(A-N)$ while the utility loss of joining when there are $i \geq f$ other members is $U(A-N+i) -$

[4] Abatement costs and environmental damages are measured in the same units. For example, integrated assessment models (e.g. DICE (Nordhaus 2008) treat abatement costs and stock-related pollution damages as output reductions.

$U(A-N+i+1-c)$. Thus the equilibrium condition under risk aversion is

$$\begin{gathered} g(p)\left[U\left(A-N+f-c\right)-U\left(A-N\right)\right]= \\ \sum_{i=f}^{N-1}\frac{(N-1)!}{i!(N-1-i)!}p^{i}\left(1-p\right)^{N-1-i}\left[U\left(A-N+i\right)-U\left(A-N+i+1-c\right)\right]. \end{gathered} \tag{2}$$

Replacing preferences by $V(U(y))$, where V is concave, and strictly concave in a subset of the interval $[A-N+f-c, A-1]$, represents an increase in risk aversion. We have the following:

Proposition 1 *(i) The introduction of risk aversion (changing the payoff from y to $U(y)$) increases the equilibrium participation probability. (ii) An increase in risk aversion increases the equilibrium participation probability.*

(Appendix A collects proofs.) The intuition for Proposition 1 is straightforward. Note that for both members and non-members, income is (weakly) increasing in the number of members. Under risk aversion, the marginal utility of income decreases with the level of income. A country loses $c-1$ units of income by joining if at least $k \geq f$ other countries have joined. The larger is k, the larger is income, and therefore the smaller is the utility loss of joining "unnecessarily" (i.e., when $k \geq f$ other countries have joined). A country gains $f-c$ by joining only if $k=f-1$, i.e. when income is low, and the marginal utility of income is relatively high. Risk aversion therefore increases the utility gain of joining when $k=f-1$ and decreases the utility loss of joining when $k \geq f$. In order for a country to remain indifferent between joining and not joining, the probability of $k \geq f$ must increase. That cumulative probability increases if and only if p, the probability that an individual country joins, increases.

Thus, in the presence of endogenous risk, an increase in risk aversion increases the equilibrium expected participation level of the IEA. To assess the magnitude of the effect of risk aversion we consider two examples here: one analytic and the other with CRRA utility.

Example: piece-wise linear utility Suppose that utility is piece-wise linear in income, with the kink at $A-N+f-c$, in order to satisfy the assumption that the

country is strictly risk averse over a subset of the interval $[A-N+f-c, A-1]$:

$$U = \left\{ \begin{array}{c} y \text{ if } y \geq A-N+f-c \\ \frac{-2w}{1-2w}(A-N-c+f) + \frac{1}{1-2w} y \text{ if } y < A-N+f-c \end{array} \right\}, \tag{3}$$

for $w \in [0, 0.5)$. With this utility function, the marginal utility of income equals 1 for $y \geq A-N+f-c$ and marginal utility equals $\frac{1}{1-2w}$ for $y < A-N+f-c$.[5] As w increases from 0 to its supremum value of 0.5, marginal utility for $y < A-N+f-c$ increases from 1 to infinity; thus, risk aversion increases with w .

The fact that utility is linear above and below the kink, $A-N+f-c$, simplifies the equilibrium condition (2). The equilibrium condition is now

$$\frac{\frac{1}{1-2w}(f-c)}{c-1} = \frac{G(p)}{g(p)}. \tag{4}$$

The numerator of the left side of equation (4) is the utility loss when the country decides not to join and there are $f-1$ members, and the denominator is the utility loss when the country decides to join and there are $k \geq f$ other members. An increase in the risk aversion parameter w increases the left hand side, due to the increase in $\frac{1}{1-2w}(f-c)$. The proofs of Propositions 1 and 2 in Hong and Karp (2012) show that the right side of equation (4), $\frac{G(p)}{g(p)}$, is increasing in p, and approaches infinity as $p \to 1$. These results imply that the equilibrium p increases in w, and approaches 1 as the risk aversion parameter approaches its supremum, $w = 0.5$. Figure 2 shows the relation between the equilibrium p and the exogenous parameter $w \in [0, 0.4995)$ for $N = 20$ and for three values of $c \in \{5.01, 5.5, 5.99\}$. For these values, equilibrium participation is 6 under pure strategies. Under risk neutrality ($w = 0$) and mixed strategies, the participation probability is less than 0.1 for all three values of c, so expected participation is less than 2. The participation probability approaches 1 (so expected participation approaches 20) as $w \to 0.5$. Hong and Karp (2012) show that, under risk neutrality, the expected membership of the mixed strategy equilibrium is

[5] The presence of the kink can be interpreted simply as risk aversion, or as the related idea of loss aversion, where positive and negative changes in the the neighborhood of a reference level, $A-N+f-c$, have asymmetric effects.

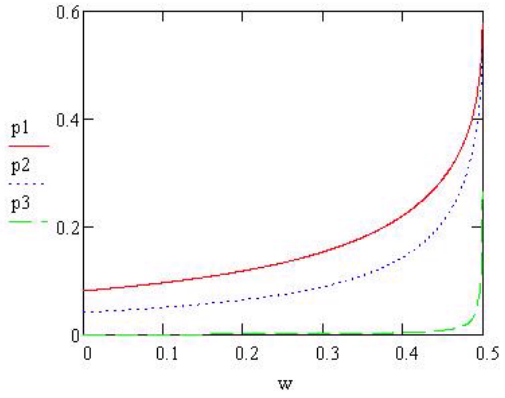

Figure 2: Relation between w and equilibrium p for $N = 20$ and $c = 5.01$ (solid), $c = 5.5$ (dotted) and $c = 5.99$ (dashed)

always lower than the corresponding membership of the pure strategy equilibrium for any $c > 2$. This result does not hold under risk aversion; for sufficiently risk averse agents, the expected membership of the mixed strategy equilibrium is higher than the corresponding membership of the pure strategy equilibrium.

Example: constant relative risk aversion (CRRA) When countries use mixed strategies and have CRRA utility

$$U(y) = \frac{y^{1-\eta}}{1-\eta} \text{ for } \eta \neq 1, \quad U(y) = \ln y \text{ for } \eta = 1, \tag{5}$$

an increase in the CRRA parameter η increases the participation probability. The equilibrium condition for p is

$$g(p)\left[\frac{(A-N+f-c)^{1-\eta}}{1-\eta} - \frac{(A-N)^{1-\eta}}{1-\eta}\right] = \sum_{i=f}^{N-1} \frac{(N-1)!}{i!(N-1-i)!} p^i (1-p)^{N-1-i} \left[\frac{(A-N+i)^{1-\eta}}{1-\eta} - \frac{(A-N+1+i-c)^{1-\eta}}{1-\eta}\right].$$

Figure 3 shows the equilibrium p as a function of η, given $N = 20$ and $c = 2.1$ for two values of $A = 10N$ (the left scale) and $A = 1.1N$ (the right scale). For $A = 10N$,

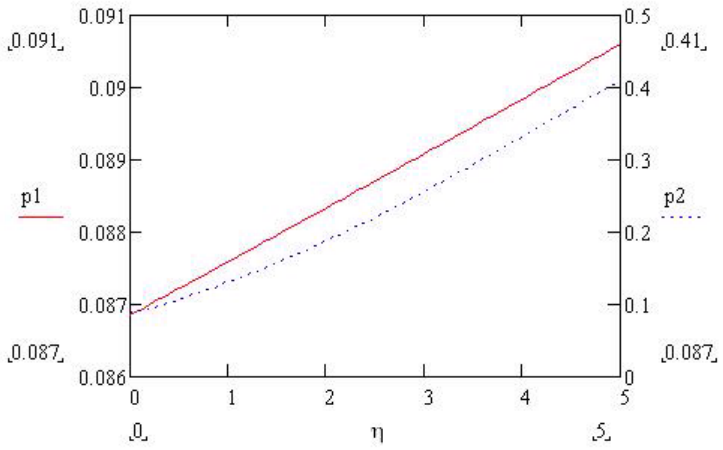

Figure 3: Relation between η and p for $N = 20$, $c = 2.1$, $A = 10N$ (left axis and the solid line $p1$) and $A = 1.1N$ (right axis and the dotted line $p2$).

environmental damage when no nation abates is only 10% of baseline income, so for all levels of participation, income is high and the marginal utility of income is nearly constant. In this case, risk aversion has little effect on the equilibrium participation probability. Membership in the pure strategy equilibrium equals 3. Expected membership varies between 1.7 and 1.9 when the CRRA parameter increases from 0 to 5.

For $A = 1.1N$, environmental damage when no nation abates is approximately 90% of baseline income. Thus, $A = 10N$ corresponds to moderate climate-related damages, and $A = 1.1N$ corresponds to catastrophic climate-related damages. With $A = 1.1N$, at low levels of participation, income is low and the marginal utility of income is high. For $A = 1.1N$, risk aversion has a significant effect on the equilibrium participation probability. Expected membership increases from 1.7 to 8.2 when CRRA parameter increases from 0 to 5. In particular, when $\eta = 2$ (a value sometimes proposed for climate policy models), expected membership is about 4, greater than the equilibrium membership under pure strategies, 3.

Comparison of examples The parameter w, which appears in the piecewise linear example, equation (3), is a measure of risk aversion in the neighborhood of the kink in the utility function. To see this, suppose that income were a random variable that takes the value $A - N + f - c - \varepsilon$ or $A - N + f - c + \varepsilon$, each with probability 0.5. Let the risk premium r equal the amount that society would pay to stabilize income at its expected value, $A - N + f - c$, leaving society with the sure income $A - N + f - c - r$. A calculation shows that $w = \frac{r}{\varepsilon}$, so w is the risk premium as a fraction of the random shock ε. Under risk neutrality ($w = 0$) the risk premium is 0, and as society becomes infinitely averse to risk ($w \to 0.5$) the risk premium approaches 0.5ε. In a general model with strictly concave utility, the supremum of $\frac{r}{\varepsilon}$ is 1.

To compare the piecewise linear and the CRRA examples, consider the case where expected income is y and actual income is $y \pm \varepsilon$, each with probability 0.5. With r defined as the amount that society would spend to stabilize income at its mean value, $\frac{r}{\varepsilon}$, is again a measure of risk aversion. The appendix shows that a second order expansion of the CRRA utility function yields

$$\frac{r}{\varepsilon} \approx \frac{1}{\eta}\left(-\frac{y}{\varepsilon} + \sqrt{\left(\frac{y}{\varepsilon}\right)^2 + \eta^2}\right). \tag{6}$$

Equation (6) and $w = \frac{r}{\varepsilon}$ (in the piece-wise linear setting) thus relate the two examples using $\frac{r}{\varepsilon}$ as a measure of risk attitude.

The ratio $\frac{y}{\varepsilon}$ provides an inverse measure of the amount of risk relative to baseline income. For the CRRA utility, as this risk becomes small ($\frac{y}{\varepsilon} \to \infty$), $\frac{r}{\varepsilon} \to 0$; as the risk approaches its maximum ($\frac{y}{\varepsilon} \to 1$), $\frac{r}{\varepsilon} \to \frac{1}{\eta}\left(-1 + \sqrt{1+\eta^2}\right)$, a quantity that varies between 0 and 1, as η increases from 0 to ∞. For example, if $\eta = 2$, then the approximation of $\frac{r}{\varepsilon}$ given by equation (6) exceeds 0.4 provided that $\varepsilon > \frac{y}{2.1}$. Thus, moderate levels of risk aversion correspond to large values of $\frac{r}{\varepsilon}$ when the relevant risk, ε, is large.

4 Exogenous risk

We consider two variants of a simple model, in order to clarify the relation between risk aversion and equilibrium membership. This relation depends on the effect of the action on the variance of income, but is unrelated to whether the action is a public good or a public bad. In both variants of the model, agents make the participation and abatement decisions before learning the realization of random abatement costs.[6] By choice of units, we set the BAU level of emissions equal to 1. We assume that marginal environmental damage is deterministic, and equal to 1. The emissions of an IEA member equals e; abatement, a, equals members' fractional reduction of their BAU emissions: $a = 1 - e$, so $0 \leq a \leq 1$. With risk averse countries and exogenous risk, abatement might take an interior value. The continuously distributed random abatement cost parameter c has support $[c_L, c_H]$, with density function $h(c)$ and expectation $\bar{c}$; c is a country's cost-benefit ratio of unilateral abatement. We adopt the following parameter restrictions:

$$(i)\ \ 1 < c_L < c_H < N;\ \ (ii)\ \ f(\overline{c}) > \overline{c}. \tag{7}$$

Inequality (7.i) implies that zero abatement is a dominant strategy for a non-member, and 100% abatement is optimal for a sufficiently large IEA, regardless of members' degree of risk aversion. Inequality (7.ii) implies that in the equilibrium under risk neutrality (where $m = f(\overline{c})$), members strictly prefer not to leave the IEA.

For Model 1, define $A^{(1)}$ as the non-random baseline income in the absence of environmental damage or emissions. The marginal benefit to a country of emissions is c, so the marginal opportunity cost of abatement is also c. Thus, in an IEA with k members, income for an IEA member equals the baseline $A^{(1)}$, minus damages associated with the non-members' BAU emissions, $N - k$, minus damages associated with the members' emissions, ke, plus the benefits of own-emisisons, ce:

$$y^{(1)} = A^{(1)} - (N - k) - ke + ce = A^{(1)} - N + ka + c(1 - a). \tag{8}$$

[6] Ulph (2004), Kolstad (2007), Kolstad and Ulph (2008) and Karp (2012) study the effect of exogenous risk and learning on equilibrium participation and welfare in an IEA.

The first equality expresses income as a function of emissions, a public bad, and the second expresses income as a function of abatement, a public good. Obviously, the characteristics of the model do not depend on which formulation we use.

For Model 2, define $A^{(2)}$ as the non-random baseline income in the absence of environmental damage or abatement. Here, each unit of abatement has the random cost c. An IEA member's income equals its baseline, $A^{(2)}$, minus damages associated with non-members, $N-k$, minus damages associated with members, ke, minus its own abatement cost, $ca = c(1-e)$:

$$y^{(2)} = A^{(2)} - (N-k) - ke - c(1-e) = A^{(2)} - N + ka - ca. \tag{9}$$

Again, the first equality here expresses the action as a public bad and the second treats it as a public good.

In Model 1, the marginal benefit of emissions (equal to the marginal opportunity cost of abatement) is a random variable. For example, the value of marginal product of an extra unit of fossil fuel may depend on the random price of an energy-intensive product. In Model 2, the marginal abatement cost is a random variable. For example, the cost of abatement may depend on the success of a technology whose properties are currently imperfectly understood. A richer model would include other random variables to account for these features, but our simpler "reduced form" models are sufficient to examine the role of risk aversion; they also nest, in a transparent manner, the familiar deterministic model. We do not regard one of these models as more plausible than the other. Consideration of both models shows that apparently minor changes in assumptions can reverse comparative static conclusions; this reversal is unrelated to whether the action is a public good or a public bad.[7]

For both models $i = 1, 2$, the IEA's policy rule is

$$a^{(i)}(k) = \arg\max_{a\in[0,1]} E_c \sum_{j\in J} U\left(y_j^{(i)}\right),$$

[7] An alternative to Model 1 treats marginal damages rather than abatement costs as random, and also implies that an increase in abatement reduces the variance of income. We can construct a different alternative to Model 2, in which abatement increases the variance of income.

where J is the set of members, with cardinality k, and j is the country index. By concavity of the IEA's maximand, $a^{(i)}(k)$ is a (single-valued) function of k. By symmetry, the IEA chooses the same action for each member, $a^{(i)}(k)$. We use the following definition.

Definition 1 *For Model $i \in \{1,2\}$, $m_a^{(i)}$ equals the minimum integer k for which $a^{(i)}(k) > 0$, and $m_b^{(i)}$ equals the minimum integer k at or above which $a^{(i)}(k) = 1$, the maximum feasible abatement level.*

The definition implies that $m_b^{(i)} - 1$ equals the maximum integer for which $a^{(i)}(k) < 1$. To establish the existence of $m_a^{(i)}$ and $m_b^{(i)}$, with $m_b^{(i)} \geq m_a^{(i)}$, note that by inequality (7.i), $a^{(i)}(k) = 1$ for all $k > c_H$, and $a^{(i)}(k) = 0$ for all $k < c_L$, for both risk averse or risk neutral agents.

With risk neutrality, $a^{(i)} = 0$ for any $k < f(\overline{c})$, and $a^{(i)} = 1$ for any $k \geq f(\overline{c})$. Recall that in general a country's incentive to join the IEA is to exercise leverage on *other* members' abatement levels. If a non-member joins, or a member leaves, its decision alters the marginal utility, to the IEA, of abatement, possibly changing the value of $a^{(i)}(k)$. Under risk neutrality, additional members above $f(\bar{c})$ have no effect on equilibrium actions, because the members are already abating at capacity; departure of a member when membership equals $f(\bar{c})$ causes the equilibrium abatement of remaining members to fall from 1 to 0, because $a^{(i)}(k) = 0$ for $k < f(\bar{c})$. Thus, in view of our tie-breaking assumption (a country that is indifferent between joining and not joining, decides to joins the IEA), the unique equilibrium under risk neutrality contains $f(\bar{c})$ members.

Matters are more complicated under risk aversion. The tie-breaking assumption implies that an IEA with $m_a^{(i)} - 2$ or fewer members is not externally stable: if one additional country joins, abatement remains at 0 and payoffs are unchanged. An IEA with $m_a^{(i)} - 1$ members is not externally stable: if one additional country joins, a positive level of abatement is optimal. Because the IEA maximizes members' welfare, the fact that it assigns a positive level of abatement, even though a zero level remains feasible, means that members' welfare is higher at the positive level of abatement. (Recall that $a^{(i)}(k)$ is a single-valued function.) An IEA with $m_b^{(i)} + 1$ or

more members is not internally stable, because a member can leave without altering the abatement levels of remaining members. The following proposition establishes existence and a basic property of equilibrium

Proposition 2 *For model $i = 1, 2$ and any level of risk aversion, there exists an equilibrium number of members, $m^* \in \left[m_a^{(i)}, m_b^{(i)}\right]$. If $m_a^{(i)} = m_b^{(i)}$, the equilibrium is unique, $m^* = m_a^{(i)}$. There is no equilibrium outside $\left[m_a^{(i)}, m_b^{(i)}\right]$. For risk neutral agents, $m_a^{(i)} = m_b^{(i)} = f(\overline{c})$.*

Proposition 2 generalizes the result that under risk neutrality, the equilibrium membership equals the "minimally successful IEA", $f(\overline{c})$. For risk neutral agents, $m_a^{(i)} = m_b^{(i)} = f(\overline{c})$, but under risk aversion it is possible that $m_b^{(i)} > m_a^{(i)}$, in which case, the equilibrium might not be unique.

The following lemma provides the ordering of the $m_a^{(i)}, m_b^{(i)}$. We use this lemma to establish and explain the subsequent proposition, which describes the effect of risk aversion on equilibrium abatement and participation.

Lemma 1 *For risk averse or risk neutral agents $m_a^{(i)}$ and $m_b^{(i)}$ satisfy*

$$f(c_L) \leq m_a^{(1)} \leq m_b^{(1)} = f(\overline{c}) = m_a^{(2)} \leq m_b^{(2)} \leq f(c_H). \tag{10}$$

Proposition 3 *(i) For any membership level k, risk aversion weakly increases abatement under Model 1 and weakly reduces abatement under Model 2, relative to abatement levels under risk neutrality. (ii) Risk aversion weakly decreases the equilibrium membership of the IEA under Model 1 and weakly increases the equilibrium membership of the IEA under Model 2.*

The essence of the comparison between Models 1 and 2 is evident from equalities (8) and (9). In Model 1, an increase in emissions (something risky) increases the variance of income; equivalently, an increase in a decreases the variance of income. In Model 2, an increase in abatement (something risky) increases the variance of income. Given risk aversion, an IEA wants to decrease the variance of members'

income. Thus, conditional on membership, higher risk aversion gives the IEA in Model 1 a greater incentive to abate; in contrast, higher risk aversion gives the IEA in Model 2 less incentive to abate. As a consequence, higher risk aversion weakly reduces a minimum critical level, $m_a^{(1)}$, in Model 1, and weakly increases a minimum critical level, $m_b^{(2)}$, in Model 2. Because the equilibrium membership always lies in the interval of $\left[m_a^{(i)}, m_b^{(i)}\right]$, and in view of the ordering in inequality (10), higher risk aversion weakly reduces equilibrium membership in Model 1 and weakly increases equilibrium membership in Model 2.

Proposition 3 describes "weak" changes of equilibrium membership. In order to provide sufficient conditions where risk aversion strictly increases or decreases equilibrium IEA membership, we use the following parameter restrictions:

$$(i)\ f(\overline{c}) > c_L + 1 \quad \text{and} \quad (ii)\ \ f(\overline{c}) + 1 > c_H > f(\overline{c}). \tag{11}$$

Inequalities (11.i) and (11.ii) imply that the range of uncertainty is non-trivial, and that the distribution of costs is not skewed "too far to the right". If the range of uncertainty is small, randomness and risk aversion are not important. If the distribution was very skewed to the right, abatement becomes unattractive for extremely risk averse agents. We have

Proposition 4 *(i) Suppose that inequality (11.i) holds. A sufficiently high risk aversion strictly reduces (relative to the risk-neutral case) membership of* ***every*** *equilibrium in Model 1. (ii) Suppose that inequality (11.ii) holds. Sufficiently high risk aversion strictly increases membership (relative to the risk-neutral case) of* ***every*** *equilibrium for Model 2.*

The logic of the proof of this proposition is the following. By Propositions 2 and 3, if $f(\overline{c})$ is not an equilibrium, then strict changes of equilibrium membership occur under risk aversion. Also, $f(\overline{c})$ is externally stable for Model 1 and internally stable for Model 2 (because $m_b^{(1)} = f(\bar{c}) = m_a^{(2)}$ by equation (10)). Therefore, for Model 1, there is a strict decrease in equilibrium membership if $f(\overline{c})$ is not internally stable, and for Model 2, there is a strict increase in equilibrium membership if $f(\overline{c})$

is not externally stable. The parameter restrictions in Proposition 4 ensure that for sufficiently high levels of risk aversion, these conditions are met.

In the interest of a complete analysis, we show that in some cases the equilibrium membership in Model 2 equals the highest possible level, $f(c_H)$, and the equilibrium membership in Model 1 equals the lowest possible level, $f(c_L)$. In some other cases, the only equilibrium in either model is $f(\overline{c})$. To establish these claims, we use the following restrictions,

$$(i)\ c_H > f(\overline{c}),\ (ii)\ f(c_H) > c_L + 1 \text{ and } (iii)\ \overline{c} > f(c_L) > c_L. \tag{12}$$

Proposition 5 *(i) Suppose that inequalities (12.i) and (12.ii) hold. For sufficiently high level of risk aversion, there exists an equilibrium with $f(c_H)$ members in Model 2 (with $f(c_H) > f(\overline{c})$). (ii) Suppose that inequality (12.iii) holds. For sufficiently high level of risk aversion, there exists an equilibrium with $f(c_L)$ members in Model 1 (with $f(c_L) < f(\overline{c})$). (iii) If inequality (11.i) does not hold, and c_L is not an integer, then for any level of risk aversion, the equilibrium membership in Model 1 is $f(\overline{c})$; if inequality (12.i) does not hold, then for any level of risk aversion, the equilibrium membership in Model 2 is $f(\overline{c})$.*

The proof of Proposition 5 is similar to that of Proposition 4 and is relegated to a referee's appendix.

In summary, for Model 1, risk aversion weakly increases abatement for given k, and as a consequence weakly decreases equilibrium participation. The reverse holds for Model 2. The reversal depends on the effect that the action has on the volatility of income, not on whether agents provide a public good or a public bad.[8]

Example: constant relative risk aversion (CRRA) Here we illustrate the role of risk aversion when the distribution of c is binary, and utility is given by

[8] In a related paper, Bramoulle and Treich (2009) consider the effect of risk aversion on emissions in a Nash equilibrium without an IEA, assuming increasing marginal environmental damage of emissions. They show that with exogenous risk, risk aversion reduces emissions in the Nash equilibrium. In their model, a reduction in emissions reduces the volatility of income. Using the arguments presented here, we can show that risk aversion has an ambiguous effect on emissions in the non-cooperative outcome, depending on how the action affects the volatility of income.

the CRRA function in equation (5). For this example, let $N = 10$, the low cost realization $c_L = 2.6$, the high cost realization, $c_H = 4.9$, and $\theta = 0.5$, the probability that the cost is low. For these values, $\overline{c} = 3.75$. Figure 4 shows how equilibrium membership, k, and total abatement level, ta, change as the CRRA parameter, η, increases from 0.1 to 10. We numerically establish uniqueness of a pure strategy equilibrium with positive abatement. The left panel of Figure 4, corresponding to Model 1, assumes $A^{(1)} = 12$; the right panel, corresponding to Model 2, assumes that $A^{(2)} = A^{(1)} + \overline{c} = 15.75$. With these values, expected income in the absence of abatement is the same in the two settings.

The green dashed curves (kn) in both panels graph the equilibrium membership and abatement level (equal to 4) under risk neutrality. In Model 1, an increase in risk aversion leads to a discrete reduction in equilibrium membership and total abatement level around $\eta = 5$; a further increase in η increases total abatement level; nevertheless, abatement level remains below 3 for $\eta > 5$. In contrast, in Model 2, an increase in risk aversion leads to a discrete increase in equilibrium membership around $\eta = 6$; total abatement is below the risk neutral level for $\eta < 6$ and above that level for $\eta > 6$.

5 Conclusion

The possibility that nations will not succeed in negotiating a (meaningful) climate agreement creates endogenous risk. Failure at the 2009 Copenhagen Conference of Parties meeting, and at various other meetings, illustrate this risk. Even if countries do reach an agreement, they face exogenous risk, arising from the inherent uncertainty about the costs and benefits of climate policy. Using mixed strategies to represent endogenous risk, we showed that increased risk aversion increases the equilibrium participation probability. Examples confirm that for high levels of risk aversion or high levels of environmental damage, countries are likely to join the IEA. These results imply that including risk aversion makes the predictions of the mixed strategy participation equilibrium less pessimistic.

In contrast, with pure strategies, the effect of risk aversion in the presence of

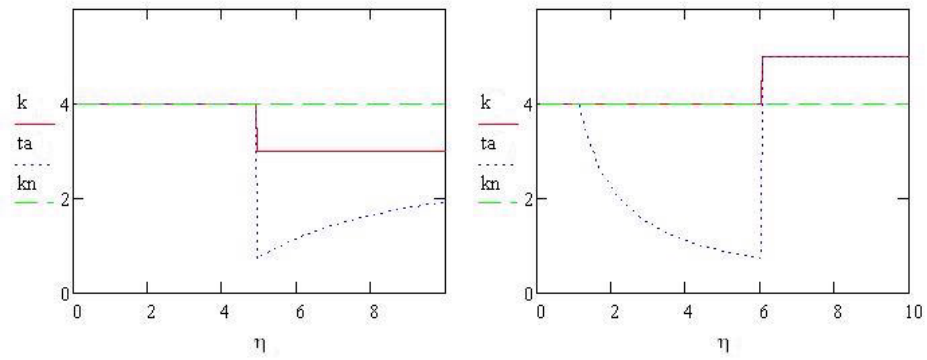

Figure 4: The left panel assumes that the baseline income without environmental damage and emissions, $A^{(1)}$, is fixed at 12 and the right panel assumes that the baseline income without environmental damage and abatement, $A^{(2)}$, is fixed at 15.75. Solid curve shows membership and dotted curve shows total abatement under risk aversion. Dashed line shows membership and abatement under risk neutrality. $N = 10$, $c_L = 2.1$, $c_H = 4.9$, $\theta = 0.5$, and η varies between 0.1 and 10.

exogenous risk has ambiguous effects on emissions and the level of participation, depending on whether abatement increases or lowers the variability of income. A simple model illustrates that in the former case, risk aversion tends to raise equilibrium participation, and in the latter case, risk aversion tends to reduce equilibrium participation.

A Appendix

Derivation of equation (6) We have

$$U(y-r) = EU = \frac{1}{2}\left(U(y+\varepsilon) + U(y-\varepsilon)\right) \tag{13}$$

A second order expansion of the left side and the right side of equation (13), evaluated at $r = 0 = \varepsilon$, yields

$$\begin{aligned} & U(y) - U'(y)r + \frac{1}{2}U''(y)\,r^2 \\ \approx\ & \frac{1}{2}\left[U(y) + U'(y)\varepsilon + \frac{1}{2}U''(y)\,\varepsilon^2 + U(y) - U'(y)\varepsilon + \frac{1}{2}U''(y)\,\varepsilon^2\right] \\ =\ & U(y) + \frac{1}{2}U''(y)\,\varepsilon^2, \end{aligned}$$

implying

$$r \approx -\frac{U''(y)\left(\varepsilon^2 - r^2\right)}{2U'(y)} = \frac{\eta}{2y}\left(\varepsilon^2 - r^2\right) \tag{14}$$

for CRRA utility. Solving equation (14) yields

$$\begin{aligned} r &\approx \frac{1}{\eta}\left(-y + \sqrt{y^2 + \eta^2\varepsilon^2}\right) \\ \frac{r}{\varepsilon} &\approx \frac{1}{\eta}\left(-\frac{y}{\varepsilon} + \sqrt{\frac{y^2}{\varepsilon^2} + \eta^2}\right). \end{aligned}$$

Proof of Propositions The proof of Proposition 1 uses the following lemma, which establishes that the functions $g(p)$ and $G(p)$ have the characteristics shown in Figure 5. In figure 5, the horizontal axis is p, the monotonic curve graphs the right side of equation (1), $(c-1)\,G(p)$, and the hump shaped curve graphs the left side, $(f-c)\,g(p)$. The equilibrium probability p under risk neutrality is determined by the intersection of these two curves.

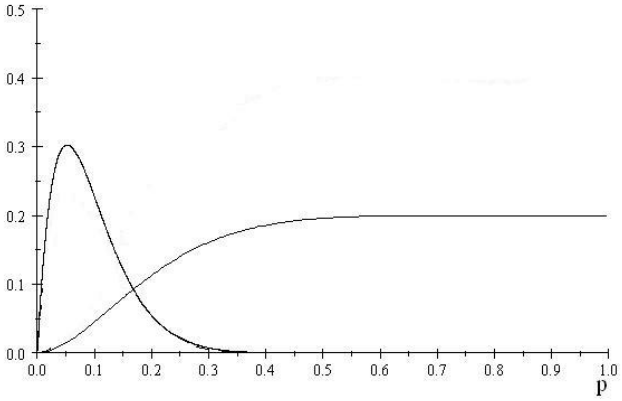

Figure 5: The monotonic curve shows $(c-1)G$ and the hump-shaped curve shows $(f-c)g$. $c = 1.2$ and $N = 20$.

Lemma 2 *We establish the following claims: (i)* $g(0) = G(0) = 0$, *and* $G(1) = 1 > g(1) = 0$; *(ii)* $g(p)$ *is single-peaked, first increasing and then decreasing in* p; *(iii)* $G(p)$ *is increasing in* p; *and (iv)* $(f-c)\,g'(0) > (c-1)\,G'(0)$.

Proof. (Lemma 2) The first claim is evident by inspection. To prove claim (ii) we take the derivative of g with respect to p to obtain

$$\frac{dg(p)}{dp} = \frac{(N-1)!}{(f-1)!\,(N-f)!} p^{f-2}\,(1-p)^{N-f-1}\,[(f-1)-(N-1)\,p]$$

Thus when $p < \frac{f-1}{N-1}$, $\frac{dg(p)}{dp} > 0$, and when $p > \frac{f-1}{N-1}$, $\frac{dg(p)}{dp} < 0$. So $g(p)$ is single-peaked, first increasing and then decreasing in p.

Claim (iii), which states $G(p)$ is increasing in p takes a bit more work. A property of binomial distribution is that $\sum_{i=0}^{N-1} \frac{(N-1)!}{i!(N-1-i)!} p^i\,(1-p)^{N-1-i} = 1$. Thus we have

$$\begin{aligned} G(p) &= \sum_{i=f}^{N-1} \frac{(N-1)!}{i!\,(N-1-i)!} p^i\,(1-p)^{N-1-i} \qquad (15) \\ &= 1 - \sum_{i=0}^{f-1} \frac{(N-1)!}{i!\,(N-1-i)!} p^i\,(1-p)^{N-1-i} \end{aligned}$$

The cumulative distribution function of a binomial distribution can be rewritten as

$$F(q;n,p) = \Pr(x \leq q) = (n-q)\frac{n!}{q!(n-q)!}\int_0^{1-p} t^{n-q-1}(1-t)^q\,dt$$

(Source: http://en.wikipedia.org/wiki/Binomial_distribution). Thus, we can rewrite equation (15) as

$$\begin{aligned} G(p) &= 1-[(N-1)-(f-1)]\frac{(N-1)!}{(f-1)!(N-f)!}\int_0^{1-p} t^{(N-1)-(f-1)-1}(1-t)^{f-1}\,dt \\ &= 1-\frac{(N-1)!}{(f-1)!(N-f-1)!}\int_0^{1-p} t^{N-f-1}(1-t)^{f-1}\,dt \end{aligned}$$

from which we easily see that $G'(p) > 0$.

We establish claim (iv), $(f-c)\,g'(0) > (c-1)\,G'(0)$, using a proof by contradiction. If $(f-c)\,g'(0) < (c-1)\,G'(0)$, then by claims (i) (ii) and (iii), $(c-1)\,G$ and $(f-c)\,g$ will either have no intersections or have two intersections at $p > 0$. But Proposition 1 of Hong and Karp (2012) shows that there exists a unique (positive) intersection that determines p under risk neutrality. Thus it is impossible that $(f-c)\,g'(0) < (c-1)\,G'(0)$. The remaining possibility is that $(f-c)\,g'(0) = (c-1)\,G'(0)$ and that the two functions are equal in the neighborhood of $p = 0$, but that again contradicts the uniqueness of the intersection. ■

Proof. (Proposition 1) Part i. Given any function U that satisfies our assumptions, we can rescale U so that

$$U(A-N+f-c) - U(A-N) = f-c \tag{16}$$

without changing preferences. Using equation (16), the left side of equation (2) is the same as the left side of the equilibrium condition under risk neutrality, equation (1). For $i \geq f$, we have

$$A-N+i > A-N+i+1-c > A-N+f-c. \tag{17}$$

Using equation (16) and the assumption that U is strictly concave in a subset of $[A-N+f-c, A-1]$, inequality (17) implies

$$U(A-N+i)-U(A-N+i+1-c)\leq c-1$$

for all $i\in[f,N-1]$ and the strict inequality holds for at least one $i\in[f,N-1]$. Consequently,

$$\begin{aligned}\textstyle\sum_{i=f}^{N-1}\frac{(N-1)!}{i!(N-1-i)!}p^{i}\left(1-p\right)^{N-1-i}\left[U\left(A-N+i\right)-U\left(A-N+i+1-c\right)\right]\\ <G\left(p\right)\left(c-1\right).\end{aligned}$$

for $p>0$. Therefore, the right side of equation (2) is less than the right side of equation (1), the equilibrium condition under risk neutrality. Meanwhile, by inspection, it is evident that the right side of equation (2) is 0 when $p=0$ and positive for any $p\geq 0$.

Using the results obtained above and Lemma 2 and referring to Figure 5, we see that the right side of equation (2) will intersect with the hump-shape curve, $g\left(p\right)\left(f-c\right)$, and the intersection is to the right of the intersection between $G\left(p\right)\left(c-1\right)$ and $g(p)\left(f-c\right)$. Hence, the equilibrium probability of participation is larger under risk aversion.

Part ii. Let V be a concave function, strictly concave over a subset of the interval $[A-N+f-c, A-1]$, so that the decision-maker with preferences V is more risk averse than the decision-maker with preferences U. Minor changes in the argument of Part i establish the result. ■

Proof. (Proposition 2) The discussion preceding the proposition establishes that the equilibrium, if it exists, must be an element of the interval $\left[m_a^{(i)}, m_b^{(i)}\right]$, so we only need to establish existence. Denote $\pi^i\left(k\right)$ and $\pi^o\left(k\right)$ as the payoff to an insider and an outsider, respectively, when there are k members, and denote $\Delta\left(k\right)\equiv\pi^i\left(k\right)-\pi^o\left(k-1\right)$, the loss to a member of leaving an IEA with k members. External stability of a candidate k requires $\Delta\left(k+1\right)<0$ and internal stability requires $\Delta\left(k\right)\geq 0$. From Definition 1 and the discussion preceding the proposition,

$\Delta\left(m_a^{(i)}\right) \geq 0$ and $\Delta\left(m_b^{(i)}+1\right) < 0$. Let $\left\{m_a^{(i)}, m_a^{(i)}+1, ...m_b^{(i)}-1, m_b^{(i)}\right\}$ be the sequence of integers between $m_a^{(i)}$ and $m_b^{(i)}$, with j'th element m^j. If $m_a^{(i)} = m_b^{(i)}$ the sequence consists of a single number that satisfies the conditions for internal and external stability, and therefore is the unique equilibrium. Now consider the case where $m_b^{(i)} > m_a^{(i)}$. Suppose, contrary to our claim, that there is no equilibrium. By definition, the first element of the sequence, $m^1 = m_a^{(i)}$ is internally stable, so if m^1 is not an equilibrium, then it must fail the test of external stability. That failure implies $\Delta(m^2) \geq 0$ (where $m^2 = m_a^{(i)} + 1$). Consequently, m^2 is internally stable. By hypothesis, m^2 is not an equilibrium, so it must fail the test of external stability, and thus $\Delta(m^3) \geq 0$. Proceeding inductively, we reach the conclusion that the final element of the sequence, $m_b^{(i)}$, satisfies $\Delta\left(m_b^{(i)}\right) \geq 0$. By hypothesis, $m_b^{(i)}$ is not an equilibrium, so it must fail the test of external stability, implying $\Delta\left(m_b^{(i)}+1\right) \geq 0$. But we have $\Delta\left(m_b^{(i)}+1\right) < 0$, contradicting our hypothesis. This completes our proof on existence. The risk neutral case is the standard model, discussed in Section 2 and in the second paragraph below Definition 1. ■

Proof. (Lemma 1) If $k < f(c_L)$, then any positive level of abatement decreases income for any possible realization of c. Therefore, the marginal utility of abatement is negative for any possible realization of c if $k < f(c_L)$. This establishes the claim that for both models $i = 1, 2$, $m_a^{(i)} \geq f(c_L)$. Similarly, if $k \geq f(c_H)$ then the marginal increase in income due to an increase in a, evaluated at $a = 1$, is positive for any possible realization of c. Therefore, for $k \geq f(c_H)$ the marginal utility of abatement, evaluated at $a = 1$, is positive for any possible realization of c. Consequently, for both models $m_b^{(i)} \leq f(c_H)$. Thus (trivially) $f(c_L) \leq m_a^{(1)}$ and $f(c_H) \geq m_b^{(2)}$.

For Model 1, given membership k, the expected marginal utility of abatement for an IEA member, evaluated at $a = 1$, is

$$\int U'\left(A^{(1)} - N + k\right)(k - c)\, h(c)\, dc = U'\left(A^{(1)} - N + k\right)(k - \overline{c}),$$

which is positive for any $k \geq \overline{c}$ (since $\overline{c}$ is a non-integer by Inequality (7.ii)) and negative for any $k < \overline{c}$. Therefore, $m_b^{(1)} = f(\overline{c})$.

For Model 2, given membership k, the expected marginal utility of abatement for a member, evaluated at $a = 0$, is

$$\int U'\left(A^{(2)} - N\right)(k - c)\, h(c)\, dc = U'\left(A^{(2)} - N\right)(k - \overline{c}),$$

which is negative for any $k < \overline{c}$ and positive for $k = f(\overline{c}) > \overline{c}$. Therefore, $m_a^{(2)} = f(\overline{c})$. ■

Proof. (Proposition 3) We use the ordering in Lemma 1 for both parts of the proof. Part (i). Under Model 1, abatement is the same under risk neutrality and risk aversion for $k < m_a^{(1)}$ and for $k \geq m_b^{(1)} = f(\overline{c})$. If $m_a^{(1)} = m_b^{(1)}$ then risk aversion has no effect on abatement in this model, for any k. However, if $m_a^{(1)} < m_b^{(1)}$, then at least for $k = m_a^{(1)}$ (and possibly also for larger k less than $m_b^{(1)}$) abatement is positive under risk aversion; but abatement is 0 under risk neutrality. Under Model 2, abatement is the same under risk neutrality and risk aversion for $k < f(\overline{c}) = m_a^{(2)}$ and for $k \geq m_b^{(2)}$. If $m_a^{(2)} = m_b^{(2)}$ then risk aversion has no effect on abatement in this model for any k. However, if $m_a^{(2)} < m_b^{(2)}$, then for $k = m_b^{(2)} - 1$, (and maybe some other k that lies in $\left[m_a^{(2)}, m_b^{(2)} - 1\right]$), abatement under risk aversion is lower than one, while abatement under risk neutrality for any $k \in \left[m_a^{(2)}, m_b^{(2)} - 1\right]$ is one.

Part (ii). Proposition 2, Lemma 1, and the fact that equilibrium membership under risk neutrality is $f(\overline{c})$, immediately imply Part (ii). ■

Proof. (Proposition 4) By Proposition 3, equilibrium membership is no larger than $f(\overline{c})$ for Model 1, and no smaller than $f(\overline{c})$ for Model 2. Therefore, if $f(\overline{c})$ is not an equilibrium, then for Model 1, risk aversion strictly reduces the membership of every equilibrium, and for Model 2, risk aversion strictly increases the membership of every equilibrium. We prove Parts (i) and (ii) by obtaining conditions under which $f(\overline{c})$ is not an equilibrium.

Part (i). By equation (10), an IEA with $f(\overline{c})$ members is externally stable for Model 1. To establish Part (i) it is necessary and sufficient to show that for

sufficiently risk averse agents, $f(\overline{c})$ is not internally stable. An outsider's income in an IEA with k members is $A^{(1)} - N + ka^{(1)}(k) + c$. An insider's income in an IEA with $f(\overline{c})$ members is $A^{(1)} - N + f(\overline{c})$, as the abatement of the insider equals 1, by $m_b^{(1)} = f(\overline{c})$. Denote $\pi^i(k)$ and $\pi^o(k)$ as the payoff to an insider and an outsider, respectively, when there are k members, and denote $\Delta(k) \equiv \pi^i(k) - \pi^o(k-1)$, the loss to a member of leaving an IEA with k members. The necessary and sufficient condition to establish that $f(\overline{c})$ is not internally stable is

$$\begin{gathered}\Delta(f(\overline{c})) = \pi^i(f(\overline{c})) - \pi^o(f(\overline{c}) - 1) < 0 \Leftrightarrow \\ U\left(A^{(1)} - N + f(\overline{c})\right) < \int U\left(A^{(1)} - N + (f(\overline{c}) - 1)\, a^{(1)}(f(\overline{c}) - 1) + c\right) h(c)\, dc.\end{gathered} \tag{18}$$

A sufficient (but not necessary) condition for the second line of Equation (18) is that the lowest possible income for the defector, which occurs when $c = c_L$, is no less than the non-stochastic income for the country that remains in the IEA:

$$\begin{gathered}A^{(1)} - N + f(\overline{c}) \leq A^{(1)} - N + (f(\overline{c}) - 1)\, a^{(1)}(f(\overline{c}) - 1) + c_L \\ \Leftrightarrow f(\overline{c}) \leq (f(\overline{c}) - 1)\, a^{(1)}(f(\overline{c}) - 1) + c_L \\ \Leftrightarrow a^{(1)}(f(\overline{c}) - 1) \geq \tfrac{f(\overline{c}) - c_L}{f(\overline{c}) - 1}.\end{gathered} \tag{19}$$

The ratio on right side lies between $(0, 1)$ by inequality (7.i). By equation (10), we know that $a^{(1)}(f(\overline{c}) - 1) < 1$. To show that for sufficiently high risk aversion there exists an interior optimal abatement level, $1 > a^{(1)}(f(\overline{c}) - 1) > 0$, satisfying the last line of inequality (19), we proceed in two steps. First, ignoring the non-negativity constraint $a^{(1)}(f(\overline{c}) - 1) > 0$, we show that an abatement level that satisfies the necessary condition for an interior optimum, monotonically increases with risk aversion and is less than 1. If this abatement level also satisfies the non-negativity constraint, then by concavity of the maximand, it is optimal. Second, we show that for sufficiently high risk aversion, we can increase the level of abatement that satisfies this first order condition, so that the levels satisfies last line of inequality (19). This level is positive and lower than one, so it must be optimal; because it

satisfies last line of inequality (19), we therefore know that $f(\overline{c})$ is not internally stable.

Step 1. Ignoring the non-negativity constraint, an interior optimal abatement level, at $k = f(\bar{c}) - 1$, satisfies the First Order Condition (FOC):

$$\int U'\left(A^{(1)} - N + (f(\overline{c}) - 1 - c)\, a + c\right)(f(\overline{c}) - 1 - c)\, h(c)\, dc = 0. \tag{20}$$

For any c such that $f(\overline{c}) - 1 > c$, U' in equation (20) is non-increasing in a and $f(\overline{c}) - 1 - c > 0$, so the integrand in equation (20) is non-increasing in a; for any c such that $f(\overline{c}) - 1 < c$, U' is non-decreasing in a and $f(\overline{c}) - 1 - c < 0$, so the integrand is also non-increasing in a. Therefore, the left side of equation (20), which represents the expected marginal utility of abatement, is non-increasing in a. (Similar statements will be made below for other first order conditions.) Meanwhile, the left side of equation (20), when evaluated at $a = 1$, equals $U'\left(A^{(1)} - N + f(\overline{c}) - 1\right)(f(\overline{c}) - 1 - \overline{c}) < 0$. Because the left side of equation (20) is non-increasing in a and is negative if $a = 1$, the abatement level that satisfies Equation (20) is less than 1.

Equation (20) can be rewritten as

$$\begin{aligned} B(a) &\equiv \int_{c_L}^{f(\overline{c})-1} U'\left(A^{(1)} - N + (f(\overline{c}) - 1 - c)\, a + c\right)(f(\overline{c}) - 1 - c)\, h(c)\, dc \\ &= \int_{f(\overline{c})-1}^{c_H} U'\left(A^{(1)} - N + (f(\overline{c}) - 1 - c)\, a + c\right)(c - f(\overline{c}) + 1)\, h(c)\, dc \equiv D(a). \end{aligned} \tag{21}$$

Inequality (11.i) ensures that $f(\overline{c}) - 1 > c_L$. The functions $B(a)$ and $D(a)$ are functionals of U, and therefore depend on the degree of risk aversion; to conserve notation, we suppress that dependence. We first consider the slopes of $B(a)$ and $D(a)$ and then explain how a change in risk aversion shifts the relative values of these functions.

The term $(f(\overline{c}) - 1 - c)$ that multiplies a in both functions B and D is positive in $B(a)$ and negative in $D(a)$. This fact and the concavity of U imply that, for fixed risk aversion and membership, $B'(a) < 0$ and $D'(a) > 0$. (Thus, the solution

to equation (21) is unique – but we already know this to be the case, as noted in the text.)

Now consider the effect of increased risk aversion on the functions (integrals) B and D. The integrands in both of these functions contain $U'\left(y^{(1)}\right)$, with $y^{(1)} = A^{(1)} - N + \left(f\left(\overline{c}\right) - 1 - c\right)a + c$. For $a < 1$, income $y^{(1)}$ increases with c, so for every c in the domain of integration of B, income and utility are lower than at any c in the domain of integration of D. Let V be any strictly concave function. Replacing preferences U with preferences $V \circ U$ implies an increase in risk aversion. With higher risk aversion, the function $U'\left(y^{(1)}\right)$ in the definitions of B and D is replaced by $V'\left(U\left(y^{(1)}\right)\right)U'\left(y^{(1)}\right)$. The new FOC is

$$\begin{aligned} B^V(a) &\equiv \int_{c_L}^{f(\overline{c})-1} V'\left(U\left(y^{(1)}\right)\right)U'\left(A^{(1)} - N + \left(f\left(\overline{c}\right) - 1 - c\right)a + c\right)\left(f\left(\overline{c}\right) - 1 - c\right)h(c)\,dc \\ &= \int_{f(\overline{c})-1}^{c_H} V'\left(U\left(y^{(1)}\right)\right)U'\left(A^{(1)} - N + \left(f\left(\overline{c}\right) - 1 - c\right)a + c\right)\left(c - f\left(\overline{c}\right) + 1\right)h(c)\,dc \equiv D^V(a), \end{aligned} \tag{22}$$

where the superscript V indicates the increased level of risk aversion. Because V is concave, and in view of our observation concerning the relative values of U over the domains of integration in the functions B and D, we know that at any two points in their respective domains of integration, the value of V' in (the risk-modified) integral B^V is greater than the value of V' in the (risk-modified) integral D^V. Consequently, replacing preferences U with preferences $V \circ U$ causes B to increase relatively more than D. (For geometric intuition, graph the downward sloping B and upward sloping D as functions of a. An increase in risk aversion could cause these graphs to shift up or down, but the comments above imply that for any a, the increase in risk aversion causes B to increase relative to D, thus increasing the point of intersection.) We therefore know that a solution to the FOC monotonically increases with risk aversion, for a given level of membership.

Step 2. To accomplish the second step, and thus complete the argument, we note that by choosing V sufficiently concave for low levels of U, and nearly linear for high levels of U, we can make the interior equilibrium $a^{(1)}\left(f\left(\overline{c}\right) - 1\right)$ arbitrarily close

to 1, and thus satisfy inequality (19). For example, we can set $V'\left(U\left(y^{(1)}\right)\right) \approx 1$ for values of $U\left(y^{(1)}\right)$ in the integrand D^V, and make $V'\left(U\left(y^{(1)}\right)\right)$ arbitrarily large for values of $U\left(y^{(1)}\right)$ in the integrand of B^V.

Part (ii). Note that $m_a^{(2)} = f(\overline{c})$, implying $a^{(2)}(f(\overline{c})) > 0 = a^{(2)}(k)$ for any $k < f(\overline{c})$. Thus, for Model 2, an IEA with $f(\overline{c})$ members is internally stable. To verify Part (ii) it is sufficient to show that $f(\overline{c})$ is not externally stable for sufficiently risk averse countries. This condition is equivalent to

$$\Delta(f(\overline{c})+1) = \pi^i(f(\overline{c})+1) - \pi^o(f(\overline{c})) \geq 0 \Leftrightarrow$$
$$\int U\left(A^{(2)} - N + (f(\overline{c})+1-c)\,a^{(2)}(f(\overline{c})+1)\right) h(c)\,dc \geq U\left(A^{(2)} - N + f(\overline{c})\,a^{(2)}(f(\overline{c}))\right). \tag{23}$$

Hence, in Model 2, risk aversion strictly increases membership for every equilibrium if and only if inequality (23) holds. A sufficient condition for inequality (23) is that the lowest possible level of income (associated with $c = c_H$) for the IEA member exceeds the defector's income:

$$\begin{aligned} A^{(2)} - N + (f(\overline{c})+1-c_H)\,a^{(2)}(f(\overline{c})+1) &\geq A^{(2)} - N + f(\overline{c})\,a^{(2)}(f(\overline{c})) \\ \Leftrightarrow (f(\overline{c})+1-c_H)\,a^{(2)}(f(\overline{c})+1) &\geq f(\overline{c})\,a^{(2)}(f(\overline{c})). \end{aligned} \tag{24}$$

Inequality (11.ii) implies, $f(\overline{c})+1-c_H > 0$; this inequality, the fact that $m_b^{(2)}$ is no greater than the smallest integer weakly above c_H, and the fact that $f(\bar{c})+1$ is an integer, imply that $f(\overline{c})+1 \geq m_b^{(2)}$; this inequality implies $a^{(2)}(f(\overline{c})+1) = 1$. Therefore, the second line of (24) holds if and only if

$$a^{(2)}(f(\overline{c})) \leq \frac{f(\overline{c})+1-c_H}{f(\overline{c})}. \tag{25}$$

Inequalities (7.i) and (11.ii) imply that the right side of this inequality lies in $(0,1)$.

We now establish that for sufficiently risk averse agents, the optimal abatement level, $a^{(2)}(f(\overline{c}))$ satisfies inequality (25). Temporarily ignoring non-negative and

less-than-one constraints, the first order condition that determines $a^{(2)}\left(f\left(\overline{c}\right)\right)$ is

$$\int U'\left(A^{(2)}-N+\left(f\left(\overline{c}\right)-c\right)a\right)\left(f\left(\overline{c}\right)-c\right)h\left(c\right)dc=0. \tag{26}$$

The left side of Equation (26) is non-increasing in a, and is positive when $a=0$ by Inequality (7.ii). Therefore, the abatement level that satisfies Equation (26) is positive. Equation (26) can be rewritten as

$$\begin{aligned} B\left(a\right) &\equiv \int_{c_L}^{f(\overline{c})} U'\left(A^{(2)}-N+\left(f\left(\overline{c}\right)-c\right)a\right)\left(f\left(\overline{c}\right)-c\right)h\left(c\right)dc \\ &= \int_{f(\overline{c})}^{c_H} U'\left(A^{(2)}-N+\left(f\left(\overline{c}\right)-c\right)a\right)\left(c-f\left(\overline{c}\right)\right)h\left(c\right)dc \equiv D\left(a\right). \end{aligned}$$

Inequality (11.ii) ensures that $c_H > f\left(\overline{c}\right)$. We abuse notation by again using the functions B and D to denote particular integrals. The rest of the proof of Part (ii) parallels the argument used to establish Part (i), so we merely sketch the steps. Income, $y^{(2)}=A^{(2)}-N+f\left(\overline{c}\right)a-ca$, increases in a over the domain of integration in B; therefore U' decreases in a over that domain. Therefore, B is a decreasing function of a. Similarly, D is an increasing function of a. At every point in the domain of integration in B, income and thus utility is higher than at any point in the domain of integration in the function D. Therefore, if we replace preferences U with preferences $V\circ U$, with V strictly concave, we have increased risk aversion, and we obtain the functions $B^V\left(a\right)$ and $D^V\left(a\right)$ together with an equation that corresponds to equation (22). Here, increased risk aversion reduces the function B relative to the function D, reducing their point of intersection. Again, by appropriately choosing the function V we ensure that the abatement level determined by the FOC, Equation (26), is no larger than $\frac{f(\overline{c})+1-c_H}{f(\overline{c})}$ (and of course less than 1). Also because this abatement level is positive as shown above, it is the optimal level $a^{(2)}\left(f\left(\overline{c}\right)\right)$. Therefore, for sufficiently risk averse agents, the optimal abatement level $a^{(2)}\left(f\left(\overline{c}\right)\right)$ satisfies Inequality 25. ■

References

Barrett, S. (1999): "A theory of full international cooperation," *Journal of Theoretical Politics*, 11(4), 519–41.

——— (2003): *Environment and Statecraft*. Oxford University Press.

Benedick, R. (2009): *Foreword to "Negotiating Environment and Science" by Richard Smith*. Resource for the Future, Washington DC.

Boucher, V., and Y. Bramoulle (2010): "Providing global public goods under uncertainty," *Journal of Public Economics*, 94, 591–603.

Bramoulle, Y., and N. Treich (2009): "Can uncertainty alleviate the commons problem?," *Journal of European Economics Association*, 7(5), 1042–1067.

Burger, N., and C. Kolstad (2009): "Voluntary public goods provision: coaltion formation and uncertainty," NBER Working Papers 15543.

Chuah, S., R. Hoffmann, and J. Larner (2011): "Escalation Bargaining: Theoretical Analysis and Experimental Test," CEDEX Discussion Paper 2011-05.

Collins, R., and K. Sherstyuk (2000): "Spatial competition with three firms: An experimental study," *Economic Inquiry*, 38, 73 – 94.

Endres, A., and C. Ohl (2003): "International Environmental Cooperation with Risk Aversion," *International Journal of Sustainable Development*, 6, 378–392.

Engelmann, D. (2003): "Risk Aversion Pays in the Class of 2 x 2 Games with No Pure Equilibrium," CERGE-EI Working Paper No. WP211.

Engelmann, D., and J. Steiner (2007): "The effects of risk preferences in mixed-strategy equilibria of 2x2 games," *Games and Economic Behavior*, 60, 381–388.

Hong, F., and L. Karp (2012): "International environmental agreements with mixed strategies and investment," *Journal of Public Economics*, 96, 685–697.

KARP, L. (2012): "The effect of learning on membership and welfare in an International Environmental Agreement," *Climatic Change*, 110, 499–505.

KARP, L., AND L. SIMON (2013): "Participation games and international environmental agreements: a non-parametric model, forthcoming," *Journal of Environmental Economics and Management*.

KOLSTAD, C. D. (2007): "Systematic uncertainty in self-enforcing international environmental agreements," *Journal of Environmental Economics and Management*, 53, 68–79.

——— (2011): *Environmental Economics*. Oxford University Press, New York, second edn.

KOLSTAD, C. D., AND A. ULPH (2008): "Learning and international environmental agreements," *Climatic Change*, 89, 125–41.

NORDHAUS, W. D. (2008): *A Question of Balance*. Yale University Press.

ULPH, A. (2004): "Stable international environmental agreements with a stock pollutant, uncertainty and learning," *Journal of Risk and Uncertainty*, 29, 53–73.

B Referee's appendix

B.1 Proof of Proposition 5

Proof. (Proposition 5) Part (i). In Model 2, an IEA with $f(c_H)$ members instructs its members to abate at capacity, so it satisfies external stability. It is also internally stable if and only if

$$\Delta\left(f\left(\overline{c}\right)\right) = \pi^{i}\left(f\left(c_H\right)\right) - \pi^{o}\left(f\left(c_H\right) - 1\right) \geq 0 \Leftrightarrow$$
$$\int U\left(A^{(2)} - N + f\left(c_H\right) - c\right) h\left(c\right) dc \geq U\left(A^{(2)} - N + \left(f\left(c_H\right) - 1\right) a^{(2)}\left(f\left(c_H\right) - 1\right)\right). \tag{27}$$

A sufficient condition for the second line of equation (27) is

$$\begin{aligned} A^{(2)} - N + f\left(c_H\right) - c_H &\geq A^{(2)} - N + \left(f\left(c_H\right) - 1\right) a^{(2)}\left(f\left(c_H\right) - 1\right) \\ \Leftrightarrow f\left(c_H\right) - c_H &\geq \left(f\left(c_H\right) - 1\right) a^{(2)}\left(f\left(c_H\right) - 1\right) \\ \Leftrightarrow a^{(2)}\left(f\left(c_H\right) - 1\right) &\leq \tfrac{f(c_H) - c_H}{f(c_H) - 1} \equiv \widehat{a}, \end{aligned} \tag{28}$$

where $\widehat{a} \in [0, 1)$. Note that expected marginal utility of abatement for a member is non-increasing in a by concavity of the IEA's maximand. If it is the case that the expected marginal utility of a member of an IEA with $f(c_H) - 1$ members, evaluated at $a = \widehat{a}$, is nonpositive, then we know that an IEA with $f(c_H) - 1$ members would set abatement no higher than $\hat{a}$; in that case, the last line of inequality (28) is satisfied. It is thus sufficient to show that for sufficiently risk averse agents, the marginal expected utility of a member of an IEA with $f(c_H) - 1$ members, evaluated

at $a = \widehat{a}$, is nonpositive, i.e.,

$$\begin{aligned}
&\int U'\left(A^{(2)} - N + \left(f\left(c_H\right) - 1 - c\right)\widehat{a}\right)\left(f\left(c_H\right) - 1 - c\right)h\left(c\right)dc \leq 0 \\
\Leftrightarrow B\left(\hat{a}\right) \equiv &\int_{c_L}^{f(c_H)-1} U'\left(A^{(2)} - N + \left(f\left(c_H\right) - 1 - c\right)\widehat{a}\right)\left(f\left(c_H\right) - 1 - c\right)h\left(c\right)dc \\
\leq &\int_{f(c_H)-1}^{c_H} U'\left(A^{(2)} - N + \left(f\left(c_H\right) - 1 - c\right)\widehat{a}\right)\left(c + 1 - f\left(c_H\right)\right)h\left(c\right)dc \equiv D\left(\hat{a}\right),
\end{aligned} \tag{29}$$

where we again abuse notation by using the functions B and D (here, with argument $\hat{a}$ rather than a). Inequality (12.ii) ensures that $f\left(c_H\right) - 1 > c_L$. The rest of the proof again parallels the arguments used in the proof of Proposition 4.i. It is straightforward to show that B is a decreasing, and D an increasing function of $\hat{a}$. At any point in the domain of integration of B, $y^{(2)}$ and U are larger than at every point in the domain of integration of D. Therefore, increasing risk aversion increases D relative to B. By choosing a sufficiently high level of risk aversion, we guarantee that inequality (29) is satisfied.

Part (ii). We show external stability first. In Model 1, an IEA with $f\left(c_L\right)$ members is externally stable if and only if

$$\begin{aligned}
&\Delta\left(f\left(c_L\right) + 1\right) = \pi^i\left(f\left(c_L\right) + 1\right) - \pi^o\left(f\left(c_L\right)\right) < 0 \Leftrightarrow \\
&\int U\left(A^{(1)} - N + \left(f\left(c_L\right) + 1 - c\right)a^{(1)}\left(f\left(c_L\right) + 1\right) + c\right)h\left(c\right)dc \\
< &\int U\left(A^{(1)} - N + f\left(c_L\right)a^{(1)}\left(f\left(c_L\right)\right) + c\right)h\left(c\right)dc.
\end{aligned} \tag{30}$$

A sufficient condition for the second line of equation (30) is for any c,

$$\begin{aligned}
A^{(1)} - N + \left(f\left(c_L\right) + 1 - c\right)a^{(1)}\left(f\left(c_L\right) + 1\right) + c &< A^{(1)} - N + f\left(c_L\right)a^{(1)}\left(f\left(c_L\right)\right) + c \\
\Leftrightarrow \left(f\left(c_L\right) + 1 - c\right)a^{(1)}\left(f\left(c_L\right) + 1\right) &< f\left(c_L\right)a^{(1)}\left(f\left(c_L\right)\right) \\
\Leftarrow a^{(1)}\left(f\left(c_L\right)\right) &\geq \tfrac{f(c_L)+1-c_L}{f(c_L)} \equiv \widetilde{a},
\end{aligned} \tag{31}$$

where $\widetilde{a} \in (0, 1)$. Note that expected marginal utility of abatement for a member is

non-increasing in a by concavity of the IEA's maximand. If it is the case that the expected marginal utility of a member of an IEA with $f(c_L)$ members, evaluated at $a = \widetilde{a}$, is nonnegative, then we know that an IEA with $f(c_L)$ members would set abatement no lower than $\widetilde{a}$; in that case, the last line of inequality (31) is satisfied. It is thus sufficient to show that for sufficiently risk averse agents, the marginal expected utility of a member of an IEA with $f(c_L)$ members, evaluated at $a = \widetilde{a}$, is nonnegative, i.e.,

$$\begin{aligned} & \int U'\left(A^{(1)} - N + \left(f(c_L) - c\right)\widetilde{a} + c\right)\left(f(c_L) - c\right) h(c)\, dc \geq 0 \\ \Leftrightarrow \quad & B(\widetilde{a}) \equiv \int_{c_L}^{f(c_L)} U'\left(A^{(1)} - N + \left(f(c_L) - c\right)\widetilde{a} + c\right)\left(f(c_L) - c\right) h(c)\, dc \\ \geq \quad & \int_{f(c_L)}^{c_H} U'\left(A^{(1)} - N + \left(f(c_L) - c\right)\widetilde{a} + c\right)\left(c - f(c_L)\right) h(c)\, dc \equiv D(\widetilde{a}), \end{aligned}$$

where we again abuse notation by using the functions B and D (here, with argument $\widetilde{a}$ rather than a). Inequality (12.iii) ensures that $f(c_L) > c_L$. The rest of the proof again parallels the arguments used in the proof of Proposition 4.i. It is straightforward to show that B is a decreasing, and D an increasing function of $\widetilde{a}$. For $\widetilde{a} < 1$, at any point in the domain of integration of B, $y^{(1)}$ and U are lower than at every point in the domain of integration of D. Therefore, increasing risk aversion reduces D relative to B. By choosing a sufficiently high level of risk aversion, we guarantee that inequality (31), which is a sufficient condition for external stability, is satisfied.

Note that in Model 1, an IEA with $f(c_L) - 1$ members makes zero abatement. By inequality (31), $a^{(1)}(f(c_L)) > 0$, then an IEA with $f(c_L)$ is also internally stable.

For Part (iii) of Proposition 5, it is obvious (by Lemma 1) that if $c_H \leq f(\overline{c})$, then $f(c_H) = f(\overline{c})$, and thus $m_b^{(2)} = f(\overline{c})$; in this case, equilibrium membership is $f(\overline{c})$ for Model 2. If $f(\overline{c}) \leq c_L + 1$, then $f(\overline{c}) = f(c_L)$ (because c_L is not an integer) and thus $m_a^{(1)} = f(\overline{c})$; here, equilibrium membership is $f(\overline{c})$ for Model 1, by Propositions 2 and 3. In these cases, risk aversion does not affect the equilibrium membership of an IEA. ■