CICERO Working Paper 2002:05

Group size effects in two repeated game models of a global climate agreement

Leif Helland

October 2002

CICERO
Center for International Climate
and Environmental Research
P.O. Box 1129 Blindern
N-0318 Oslo, Norway
Phone: +47 22 85 87 50
Fax: +47 22 85 87 51
E-mail: admin@cicero.uio.no
Web: www.cicero.uio.no

CICERO Senter for klimaforskning
P.B. 1129 Blindern, 0318 Oslo
Telefon: 22 85 87 50
Faks: 22 85 87 51
E-post: admin@cicero.uio.no
Nett: www.cicero.uio.no

Tittel: Group size effects in two repeated game models of a global climate agreement

Forfatter(e): Leif Helland

CICERO Working Paper 2002:05
16 sider

Finansieringskilde: Norges forskningsråd

Prosjekt: Håndheving, verifikasjon og design av klimaavtaler

Prosjektleder: Jon Hovi

Kvalitetsansvarlig: Jon Hovi

Nøkkelord: Håndheving,

Sammendrag:. Hvor store totale reduksjoner i utslippet av klimagasser kan en forvente i en global klimaavtale? Innenfor spillteorien er det formulert potensielle svar på dette spørsmålet. Dette arbeidsnotatet inneholder en kritisk diskusjon av to (sentrale) spillmodeller som gir pessimistiske svar på spørsmålet. Begge modeller tar utgangspunkt i et uendelig gjentatt, n-person fangens dilemma spill. Den første modellen er en full informasjonsmodell og krever at likevektene må være svakt reforhandlingssikre. Dette leder til en (kanskje kontraintuitiv) konklusjon om at en avtale som kan gi høy nytte for gruppen, vil gi lavere totale reduksjoner enn tilfellet er for en avtale som kan gi lav nytte for gruppen. Den andre modellen forutsetter ufullkommen offentlig informasjon og krever at likevekten må hvile på terskelutløsende strategier. Hovedkonklusjonene (som er mer intuitive) er to: Totale reduksjoner tiltar med forbedret verifikasjonsteknikk, for en gitt spillergruppe. Men det er også slik at totale reduksjoner tiltar med økende spillergruppe, for gitt verifikasjonsteknologi. Empiriske implikasjoner fra de to modellene identifiseres, og det argumenteres for at disse bør konfronteres med eksperimentelt genererte data, slik at en kan diskriminere mellom modellene på bakgrunn av deres prediksjonskraft. En grunn for dette er at det ikke finnes historiske data på avtalte reduksjoner i utslipp av klimagasser, siden ingen slik avtale enda har trådt i kraft.

Språk: Engelsk

Title: Group size effects in two repeated game models of a global climate agreement

Author(s): Leif Helland

CICERO Working Paper 2002:05
16 pages

Financed by: Research Council of Norway

Project: Compliance, verification, and design of climate treaties

Project manager: Jon Hovi

Quality manager: Jon Hovi

Keywords: Compliance,

Abstract: What levels of total abatement can one hope for in a global climate agreement? Some potential answers to this question are provided by game theory. This working paper contains a critical discussion of two (prominent) game models that answer the question quite pessimistically. Both models take the n-person, infinitely repeated prisoner's dilemma game as their point of departure. The first model is a full information model and utilizes the notion of a weakly renegotiation proof equilibrium. This results in the (maybe counterintuitive) prediction that an agreement that can provide high utility to the group will attract less total abatement than an agreement that can only provide low utility to the group. The second model assumes imperfect public information and utilizes the notion of a trigger level equilibrium. This results in the (more intuitive) prediction that the level of total abatements will increase with improved verification techniques, for a given player set. Still the level of total abatements decrease with an increasing player set, for a given verification technique. Empirical implications of the two models are identified, and it is argued that one should confront these with experimentally generated data in order to discriminate between the models. One reason for this is that historical data on abatement efforts in a global climate agreement do not exist since no such agreement has entered into force yet.

Language of report: English

Rapporten kan bestilles fra:
CICERO Senter for klimaforskning
P.B. 1129 Blindern
0318 Oslo

Eller lastes ned fra:
http://www.cicero.uio.no

The report may be ordered from:
CICERO (Center for International Climate and Environmental Research – Oslo)
PO Box 1129 Blindern
0318 Oslo, NORWAY

Or be downloaded from:
http://www.cicero.uio.no

Contents

Acknowledgements

The author gratefully acknowledges valuable comments from Scott Barrett and Jon Hovi, as well as from the participants in a workshop on compliance and verification in Oslo (31.05 – 01.06 2002). The usual disclaimer applies.

1 Introduction

How much cooperation can one reasonably expect from an interstate agreement to reduce emissions of greenhouse gases to the atmosphere? Can one hope to increase the amount of such cooperation through mindful design of the agreement? If so, which design principle is best suited to enhance cooperation? Finding plausible answers to such questions is clearly an important endeavour. One common and influential way of seeking insight is to employ the tools of game theory, a stylized theory of strategic interaction between consciously goal seeking actors.

This chapter contains a critical discussion of some of the answers provided by two game theoretic models that figure prominently in the literature on global warming. Both models produce pessimistic predictions. To some extent the design principles implied seem counterintuitive. Furthermore, the predictions are largely untested against data. For this reason the predictions and design proposals should be viewed as preliminary, but highly interesting, hypotheses. In line with common norms of scientific practise, we ought to require that predictions stand their ground in empirical confrontations before regarding them as valid explanations, giving rise to attractive design proposals. In the last part of the chapter, some thoughts on how we may design suitable empirical tests of the predictions are presented.

1.1 The framework, and two repeated game models

Incentive problems relating to the decentralized provision of collective goods are frequently analyzed as games of strategic interaction between 'players' having 'prisoner's dilemma' payoffs. A collective good is a good that every player benefits from once it is provided, no matter whether the player contributed to the provision of the good or not. Since the costs of provision are borne by individual players, and since the benefit of a single provision is taken to be marginal, a free rider incentive is created. Every player prefers that the other players provide the good and that they can abstain from costly contributions altogether.

Reduced emissions of greenhouse gases are commonly viewed as *the* epitome of a collective good. Reducing the probability of adverse consequences in the future require individual states to undertake costly emission reductions. The state system is truly decentralized, due to the principle of sovereignty. Any emission reductions undertaken locally will lessen the probability of adverse consequences on a *global* scale because of the way greenhouse gases mix in the atmosphere. Even though a particular set of reductions may have distributive effects, with the best of contemporary knowledge these remain fundamentally uncertain.[1] The problem is therefore commonly simplified by assuming that the benefits of a particular reduction in the probability of adverse consequences are spread equally among the states. This simplification is used in the following discussion.

The maximal reduction that any *single* state is able to undertake is assumed to have a fairly small impact on total global warming. Since emission reductions are costly, each state faces a free rider incentive. A similar incentive faces any signatory that happens to have entered into a costly climate agreement. In this case the behavioural expression is non-compliance with agreed upon (and costly) emission reductions. Such incentives are magnified to the extent that unilateral emission reductions reduce the competitiveness of the state in world markets.

A stylized example of the strategic incentives facing a single state in a situation like this is provided in figure 1. The horizontal axis gives the number of *other* states *contributing* to the

[1]On this it is interesting to compare the accurate prophecy on the knowledge situation 10 years ahead found in Schelling 1992:2-3, with the statements of the knowledge situation found in UNEP2001.

provision of the good, n. The maximal number of other contributors is N-1. The vertical axis gives the payoffs for the state in question, say state I, as a function of n. For the sake of simplicity, we assume that each state has a binary choice: either to contribute a given amount to the provision of the good, or abstain from contributions all together.[2] The function $G_i(n)$ give the payoffs of abstaining while the function $F_i(n)$ give the payoffs of contributing, both as functions of the number of other contributors. The parameter c designates the individual cost of a contribution. A prisoner's dilemma – like that in figure 1 – is characterised by two relationships. First, for any number of contributors, it is always best for each player to abstain from contributions ($G_i(n)>F_i(n)$ for all n and for any $i \in \{1,..,N\}$). Second, universal cooperation is preferred to universal non-cooperation by any player ($F_i(N)>G_i(0)$ for any player $i \in \{1,..,N\}$). The dilemma consists in the following. Abstaining is a dominant strategy for any player – that is, a strategy that every rational player would choose since it is a best response no matter what the other players choose. At the same time the resulting equilibrium is inefficient in the Pareto sense; that is, every player prefers the (non-equilibrium) outcome where they all contribute, to the (equilibrium) outcome where no player contributes. At least such conclusions hold if the game is played a finite number of periods under conditions of full information.[3]

In other words, the prediction of the finitely repeated game is that even in a situation where every signatory to a climate agreement prefers all signatories to honour their obligations and reduce the emissions of greenhouse gases rather than cheat on their obligations by non-compliance, no signatory will heed its obligation. Foreseeing this, no rational party will want to enter a climate agreement in the first place. The pessimistic behavioural prediction of the finite horizon game is indicated by the black disc in figure 1.

If the interaction is modelled as an 'infinitely' repeated game, the behavioural predictions can be less gloomy than in the case where the game is played a finite (and known) number of periods. By 'infinitely' we understand that the interaction has a constant periodic probability of continuing for yet another period, where the periods are taken to be of equal length. Specifically, if the exact length of the repeated game is unknown in this sense, and the states are patient, it cannot be rule out that some provision of the good will be an equilibrium outcome. Arguably, an 'open horizon' game is more plausible as far as global warming goes. A 'last period' in the global warming game is really only conceivable insofar as major breakthroughs in clean technologies can be envisioned (examples that spring to mind are fission energy and hydrogen engines). It is of course difficult, or even impossible, to predict a precise arrival date for such breakthroughs. The model-notion of a constant continuation probability may seem passable as an analytical substitute for the possibility of a technological breakthrough.

In what follows, attention is limited to the open horizon game. How much provision one may expect in such a game varies with the details of the model. The first model we consider is from Barrett (1999). Barrett's model leads to particularly distressing predictions. Specifically it predicts that in a given group of rational states, (1) only climate agreements of *little* value to the group will be implemented by the full group of states, and (2) climate agreements of *some*

[2]This assumption is captured in the figure by letting the distance between states on the horizontal axis be equal.

[3]The reason for this goes as follows. In a finitely repeated game, non-compliance must be a dominant strategy in the ultimate stage game. After the ultimate stage game there is now a future in which one may be punished for non-cooperation. Knowing this, players in the penultimate stage game will realise that non-compliance is a dominant strategy here also. This logic carries over all the way to the first stage game in the finitely repeated game, and is commonly referred to as the unravelling problem.

value to the group will only be implemented by a small number of the states, rendering such cooperation of little value to the full group of states.

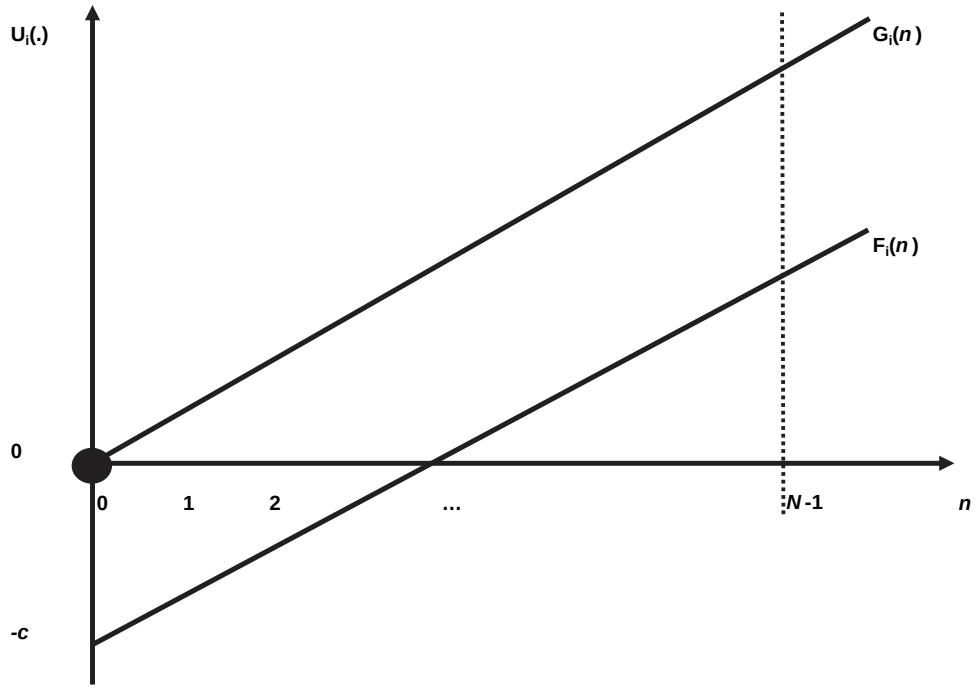

Figure 1. A prisoner's dilemma game

For instance, Barrett argues that the achievements of the Montreal agreement on ozone depletion could carry a large number of signatories partly because it achieved little more than a modernization of technology, which would have been modernized in any case. In this sense the agreement was of little value to the group of signatories, and the group of signatories could be large. In spite of this, the agreement had to be bolstered by side payments and trade restrictions.

In the global warming case, abatement costs are far from negligible. Accordingly, one might expect to see few signatories to an agreement on global warming. Alternatively, if the number of signatories eventually turns out to be substantial, one would expect the terms agreed upon to render the agreement virtually valueless to the signatories. If we choose to tell the history of the Bonn meeting through Barrett's perspective, two features stand out: The US decided not to become a signatory, and the terms of the Bonn agreement were substantially watered down by allowing sinks to count toward meeting reductions targets (see Torvanger 2001 for details).

In my view, a counterintuitive implication of Barrett's model is that it is possible to make an agreement easier to comply with by rendering it less efficient (or otherwise more costly). In this way, participation in an agreement on global warming may be broadened by making it *less* attractive.

Of course, in Barrett's model cost and benefits are givens: The researcher is allowed to manipulate with such entities in order to obtain comparative statics. Players, however, are not

allowed to choose costs or benefits. In reality, however, parties must coordinate their beliefs on the terms of a treaty (by negotiation or otherwise) before the question of signing it or not has any meaning. There are many ways in which costs and benefits can be, and are, manipulated. It suffices to mention the choice of instruments of implementation (for instance uniform emission reductions or taxes on emission); the question of whether trade in quotas should be allowed, and if so how such a market ought to be designed; and the question of whether credits for sinks should be allowed or not.

In my opinion this is precisely why Barrett's comparative static results are interesting. Thus, we should pay careful attention to the prediction that parties can coordinate on a less efficient treaty in order to enlarge the group of signatories. To me this facet of the model is contrary to the usual intuition of economists, namely that more efficient agreements will facilitate compliance.

Like any model, Barrett's model rests on a number of restrictive assumptions. Section 2.3 contains a critical discussion of the model assumptions, the robustness of the model predictions and their (lack of) intuitive appeal.

In section 3.1 another variant of the open horizon game is presented, namely that of Green and Porter (1984). It deviates from the one discussed in section 2 in that states have *imperfect public information*: This means that no state can observe another's actions directly, but that every state can observe a *random variable* that may change for two reasons: (1) (Non-) compliance with the climate agreement. (2) Natural fluctuations not due to (non-) compliance. The probability distribution over fluctuations is taken to be commonly known.[4] This model was originally formulated to analyze cartel agreements in economics. Since uncertainty about natural fluctuations is endemic in the global warming question, such a model formulation may seem promising. Section 3.2 discusses the model's applicability to climate agreements. In particular, the question of *what kinds* of random variables lend themselves to the type of monitoring envisaged in the Green-Porter model is addressed. In Barrett's model, monitoring and verification can (of course) have no place, since information is assumed to be (almost) perfect and complete (for convenience we refer to this as 'a full information environment').[5]

The Green-Porter model shares a prediction with Barrett's model, namely that less cooperation is more likely in a large group of signatories than in a small group. The reasons given for this, however, are quite different in the two models. In the Green-Porter model, states cooperate if the state-variable does not exceed an agreed upon level. The marginal impact of non-compliance on the random variable, however, is smaller the larger the group, so the incentives to comply are weakened in a large group. In Barrett's 1999 model, less cooperation can be expected in a large group because the particular collective rationality requirement imposed forbids it. At the same time it is noted that if one imposes the same kind of collective rationality in the Green-Porter model, cooperation also decreases in this model.[6]

In the Green-Porter model, verification may play a role because the environment is not one of full information. And interestingly, it turns out that more cooperation can be extracted by reducing the spread of the random-variable in this model or by improving verification

[4] 'Commonly known' meaning that every player knows it, knows that every player knows it, that every player know that every other player knows it, and so on.

[5] Perfect information means that the players have no uncertainty as to where in the game they are when they take actions. In the game at hand the information is almost perfect since players do not know each others choice of actions when actions are taken in the present period, but they do know the actions taken in all subsequent periods. Complete information means that there is no uncertainty (of any order) as far as the other players 'types' are concerned. Types refer to payoff-relevant characteristics of the players, such as utility over outcomes and beliefs about the relevant parameters of the game.

[6] See van Damme (1989), Kong-Pin (1995).

techniques. Thus, the model singles out verification and monitoring as variables that the design of a climate agreement ought to focus on. This is elaborated on in section 3.3.[7]

According to the standard view on methodology, model predictions should be evaluated by confrontations with real world data. Unfortunately, but for obvious reasons, no data on (non)compliance with a global climate agreement exists. In section 4 the possibility of testing competing predictions of the two models in a laboratory setting is discussed. Sketches for the design of such experiments are provided, and references to central results from existing experiments on decentralized provision of public goods are noted. Section 5 summarizes and concludes the chapter.

2 Compliance with full information

Provided that the parties are patient enough (if the discount factor is high enough), almost any conceivable pattern of behaviour may be constructed as a sub-game perfect equilibrium in the open horizon game.

In a sub-game perfect equilibrium no party has any incentive to deviate unilaterally from the prescribed behaviour in any sub game, on or off the equilibrium path of the play. In the infinite horizon game each period constitutes a separate sub game. Consider as an example an equilibrium in which every player plays the famous 'grim trigger' strategy. First, this strategy tells a player to comply with the terms of the agreement, as long as no player has deviated from those terms (on the equilibrium path). Second, the 'grim trigger' tells a player never to comply with the agreement again if any deviation from the agreed upon terms is ever observed (off the equilibrium path). It can be shown that even fairly impatient players may establish a threat structure that ensures universal compliance, if they all abide by the grim trigger.

If every player sticks to «grim trigger» each will obtain a discounted payoff of $F_i(n)/(1-\delta)$. A single deviation by player *i* will net him a payoff of $G_i(n-1)$ in the sub game when he deviates. Thereafter he will get a discounted payoff of $\delta G_i(0)/(1-\delta)$, since the agreement is suspended for ever by every player after a deviation has been observed. Thus if $F_i(n)/(1-\delta) \geq G_i(n-1) + \delta G_i(0)/(1-\delta)$ it does not pay to deviate from the agreed upon terms. Without loss of generality we may normalize $G_i(0) = 0$ (as in figure 1). Now it is easily seen that a single deviation does not pay the player if $\left[G_i(n-1) - F_i(n)\right]/G_i(n-1) \leq \delta$. If the condition is fulfilled, grim trigger is a best reply on the equilibrium path.

It is also easy to see that no player can gain by not carrying out the proscribed punishment (of suspending the agreement for ever) once a deviation has been observed. Given that every other player carries out the punishments proscribed by grim trigger, player *i* will get a periodic payoff of $G_i(0) = 0$ if he too carry out the punishment. If player *i* continue to fulfil his obligations he will get a periodic payoff of $F_i(0) < 0$ (by definition of the prisoners dilemma game). It is therefore in a players interest to carry out the equilibrium threat if called upon to do so. The threat is therefore said to be credible, and grim trigger is accordingly a best reply off the equilibrium path.

[7] This discussion relates to the model presented in chapter 7 of the book.

A simple numerical example may clarify. Consider a situation where the agreement is worth one unit of utility for each additional player that complies, and where the cost of compliance is two units of utility. In an agreement where 101 players play the «grim trigger», unilateral deviations does not pay off as long as $\left[(100*1)-(101*1-2)\right]/100 = 1/100 \leq \delta$. A discount factor of 1/100 means that postponing the benefits of reduced emission for one period, is worth only 1/100 of having these benefits right away. Such a discount factor can hardly be said to characterize particularly patient players. The numerical example thus illustrates that even very impatient players may (in theory) establish a treat structure that ensures unanimous compliance to the agreed upon emission reductions as sub game perfect equilibrium.

If the discount factor is high, milder threats than a grim trigger will insure that no player has any incentive to deviate unilaterally. Such threats may take more or less peculiar forms, for instance a threat to postpone the agreement for $t<\infty$ number of periods; a threat to never comply in odd numbered periods ever again; a threat to suspend the agreement for t number of periods followed by compliance for t' number of periods, x number of times after a deviation, and so on. This phenomenon is referred to as 'the folk theorem'. It renders the model without empirical cutting power, since any conceivable pattern of behaviour can be constructed as a sub-game perfect equilibrium.

2.1 Barrett's model

Barrett refers to a sub-game perfect equilibrium as individually rational. A few years ago an agreement that rested on such an equilibrium was commonly referred to as a self-enforcing agreement because no party has an individual incentive to deviate unilaterally from the specified terms (on or off the equilibrium path) as long as the agreement rests on strategies that constitute a sub-game perfect equilibrium.

Still, there is something unsatisfactory in equating the term 'self enforcing' with the behaviour dictated by a strategy like 'grim trigger'. Once play is brought off the equilibrium path by a single deviation, it is in every player's interest that they all – collectively – abandon their punishments and restart cooperation. Knowing this, a player will understand that if he deviates, claims that the deviation was a one-of-a-kind incident never to be repeated, and proposes that the deviation be pardoned, then rational players will pardon, since if they collectively do so they will each increase their individual payoff. But then every player will want to deviate and be pardoned in every period, the logic of the threat structure unravels, and it seems curious to refer to the agreement as 'self-enforcing'.

To be truly self-enforcing Barrett demands that an agreement should not only be individually rational (in the sense of sub-game perfect), but also collectively rational, in the sense that the players cannot profit individually by deviating collectively from the equilibrium threats of punishing transgressions of the terms. The strategy Barrett considers is a close relative of the famous 'tit-for-tat' strategy, namely 'getting even'. While tit-for-tat is not sub-game perfect, getting even is. Like tit-for-tat, getting even uses a far milder threat of punishments than does the grim trigger. Getting even says that a player contributes unless he has contributed more frequently than any other player in previous rounds. If he has contributed more frequently in previous rounds, he gets even by not contributing while the others contribute. In this sense getting even matches the punishment to the harm done.

To see what collective rationality means in the infinite horizon game we utilize some simple notation. Following Barrett we close the model by letting the payoff functions be linear in the number of other players cooperating. More precisely we let $F(n) = f \cdot n - c$ and $G(n) = g \cdot n$, where c is the cost of contributing. As before, we normalize $G(0) = 0$ without loss of generality. It is easily shown that the strategy vector where every player

follows 'getting even' is a sub-game perfect equilibrium, provided $c > g \geq f > 0$ (see Barrett 1999:531 for the details). That this strategy vector is sub-game perfect, however, is not hot news, considering the above-mentioned folk theorem.

We therefore turn to the condition for the strategy vector where all players stick to 'getting even' to be renegotiation proof. Assuming discount factors arbitrarily close to one, the condition is not difficult to get a grip on. If a player deviates, each of the other signatories will net a payoff of $f \cdot (n-1) - c$ in that period. In the period immediately following the deviation, the deviating player (in accordance with the dictums of 'getting even') pays penance by being the sole provider of the good. This nets each of the other players $g \cdot (1)$. Carrying out the threat dictated by 'getting even' thus nets each of the punishing players a total of $f \cdot (n-1) - c + g$. If the non-deviating players collectively decide to forget the single deviation (in contradiction with the dictums of 'getting even'), their payoffs will be $f \cdot (n-1) - c + f \cdot n - c$. Thus, it is not individually profitable to forget collectively if $f \cdot (n-1) - c + g \geq f \cdot (n-1) - c + f \cdot n - c$, or in other words if $(g + c) / f \geq n$.

It is this last condition that give rise to the central - and in my opinion counterintuitive - predictions of Barrett's model: first, that a valuable climate agreement can only be self-enforcing for a small number of signatories; and second, that a climate agreement which is valuable for the signatories can only be self-enforcing if the number of signatories to the agreement is small.

For the sake of building intuition, consider the same numerical example as above. Let each contribution from the other players increase a players benefit with one unit of utility, while the cost of contribution is two units of utility for every player. Using Barrett's condition we find that an agreement will be self enforcing if, and only if, $(1 + 2) / 1 \geq n$.

In words: if the number of signatories to the climate agreement is less than or equal to three it will be self enforcing, otherwise not. The potential number of signatories in the example was 101. So; if we want to be sure that the 101 players affected are able and willing to heed their obligations to reduce emissions of green house gases, only three of them should become signatories to the climate agreement, given the benefits and costs of the example. The remaining 98 states should be allowed to free ride, so as to assure that the agreement is self enforcing.

If we wish to include more signatories to a self enforcing climate agreement, however, we may do so. But this comes at a price. For instance we may make it more costly for the states to abate. This can be achieved in a number of ways. One way is to prohibit such measures that economists usually want to build climate agreements on, like efficient emission taxation, markets for tradable emission quotas, and joint implementation mechanism. For instance, by rising the cost to five units of utility the agreement can carry six signatories and still be self enforcing; and by rising the cost to 100 units of utility even the full group of 101 states can be signatories to a self enforcing climate agreement.

2.2 Relevance

Barrett's model rests on a number of restrictive assumptions. Among them are the following:

(i) All countries are identical.

(ii) Interaction takes place under (almost) perfect and complete information.

(iii) Cooperation exhibits constant or increasing returns (since $G_i(n)$ is assumed to increase at a slower rate than $F_i(n)$).

(iv) A uniform emission quota is the only instrument available.

(v) Abatement levels from previous periods are observed noiselessly at the beginning of every period, and without costs.

(vi) Punishments can be carried out with full force immediately after observing abatement levels.

(vii) Only internal threats (strategies of reciprocity) are considered (the only sanction allowed is reacting to non-compliance by reducing one's own abatement efforts).

(viii) Cost functions are independent (which, for instance, means that interaction effects via world markets are not taken into account).

(ix) Choice of abatement levels are binary (either socially efficient or no abatement)

Barrett's pessimistic predictions may consequently be unwarranted: Relaxing one or more of the above assumptions may after all result in more optimistic conclusions. What then, do we know about the robustness of the model conclusions to relaxations in the assumptions? The answer seems to be 'not a lot'. To the best of my knowledge the only assumption relaxed is (iv). This is done by Finus (2001:274-279) and by Finus and Rundshagen (1998), which allow for endogenous choice between uniform emission reduction quotas and uniform emission taxes. The conclusions arrived at do not differ qualitatively from the ones reached in Barrett's 1999 model. So in this respect Barrett's model seems robust.

Models of climate cooperation building on quasi-dynamic processes of conjectural adaptation have been forwarded (Barrett 1994, Bauer 1992, Carraro and Siniscalco 1992,1993, Hoel 1992, cf. Finus and Rundshagen 1998:146-7, Finus 2001 and Wagner 2001:378-88 for discussions). In such models, actions are not taken in real time. Rather the dynamics are thought of as a cognitive process taking place in the minds of the players. Actions are taken once and only once, and such models should therefore be viewed as essentially static. Consequently the equilibria of such models may deviate from the equilibria in a truly dynamic model, so 'the adjustment process itself may not be an equilibrium of the repeated game where players know that they face each other repeatedly' (Fudenberg and Tirole 1991:26). In addition to this, players' beliefs are given exogenously by their reaction functions in the quasi-dynamic models. This may in turn lead players to entertain updated beliefs that are not rational in the usual sense of the word (that is, their beliefs may not follow from the use of Bayes' rule where this is possible).[8]

Bearing these serious limitations in mind, we may nonetheless note that if collective rationality is imposed, the conjectural adaptation models tends to lead to the same pessimistic conclusion as Barrett's 1999 model does. Furthermore, some of the above assumptions have been relaxed in the static models, without producing qualitatively more optimistic conclusions.[9] Specifically, the pessimistic conclusions holds when (i) is relaxed and countries are allowed to differ in size or in payoff functions; when (iii) is relaxed and cooperation is allowed to exhibit decreasing returns; when (vii) is relaxed and issue linkage is allowed (for instance by way of trade sanctions, or more generally by way of 'side payments').[10]

[8]See Rasmusen (1994:312-13) for an elaboration on this point. See also Finus and Rundshagen (1998:146-47), who argue that quasi dynamic models are unrealistic in a specific sense.

[9]The adaptive thought experiment ends in a Nash equilibrium for the corresponding simultaneous move game. Examples of such thought experiments are found in Cournout and Stackleberg's convergence.

[10]A good overview of the results obtained within the framework of static models is given by Finus 2001:219-57.

2.3 Discussion

As noted, Barrett's 1999 model rests on a number of highly restrictive assumptions, which makes the resemblance between it and real climate negotiations fairly vague. Little is known about how robust the predicted difficulties in establishing and maintaining an ambitious climate agreement are with respect to relaxations in such assumptions. What we 'know' about this stems from models that do not properly take time into account. Such 'knowledge' should be handled with care, since it is fundamental to social life that actions are actually taken in real time (not by abstract contemplation).

Given the set of assumptions, Barrett's model results in (what I find to be) counterintuitive conclusions. Being counterintuitive, we must wonder whether *real* people placed in an environment mimicking the one in the model will actually behave as predicted. On this point, laboratory experiments may turn out to provide guidance. If carefully constructed laboratory experiments fail to support the model implications, we may suspect that there is something fundamentally wrong with Barrett's model. That would be less trivial – and far more interesting - than merely pointing out that his model (like all models) rests on a set of restrictive assumptions: For even a model resting on blatantly restrictive assumptions may well provide the perfect stepping stone to more refined hypotheses at a later stage.

In my view, two assumptions stand out as potentially being 'fundamentally wrong'. First, in open-horizon games the concept of collective rationality is ambiguous even in the two player case. Several definitions of renegotiation proof equilibria have been forwarded for this class of games, but the different definitions do not necessarily lead to the same behavioural predictions (cf. the discussion in Fudenberg and Tirole 1991:179-182). Except for Barrett's own work, attempts to clarify what collective rationality means in infinite horizon games with *more* than two players are lacking. Barrett uses one particular definition of collective rationality (based on Farell and Maskin 1989) in his 1999 paper, but does not discuss alternative definitions.

In a more recent paper, however, (Barrett 2002) strengthens the notion of collective rationality. This turns out to support additional cooperation in a larger group of signatories. Qualitatively, however, the implications resemble the ones of the 1999 paper – particularly '[f]or any given problem an SCR [strongly collectively rational] treaty is unique and typically incomplete; some countries cooperate and some free ride' (2002:3). Though the modification of the collective rationality concept is interesting, it will not be dealt further with here.

Second, we should note that there is a deep distributional conflict between signatories and non-signatories in Barrett's model. Barrett's model provides no guidance as to the selection between the equilibria the model gives rise to. We may suspect that the way we model this selection process can have a profound impact on the conclusions we are able to draw about the kind of climate agreements that will be entered into and complied with. In any situation involving real people - for instance in controlled experiments - this selection problem is very real, and will have to be solved. Also on this point more work on the models is needed for us to form firm conjectures.

To see this second point more clearly, we continue the numerical example from above. There are 101 potential signatories but only 3 signatories are allowed in the self enforcing agreement. Thus the 3 signatories each obtain a net periodic payoff of $f \cdot n - c = 1 * 3 - 2 = 1$ while the corresponding payoff for each of the 98 non signatories is $g \cdot n = 1 * 3 = 3$. Every player evidently prefers to become a non-signatory. With 101 players, however, there exists (by the binomial coefficient) 166 650 ways to form a group of 3 actual signatories. Or in other words; we encounter an overwhelming number of renegotiation proof equilibria, and they all give rise to deep distributional conflicts.

3 Compliance with imperfect public information

All this aside, the most unsatisfactory assumption made by Barrett may well be that information is (almost) perfect and complete. A full information environment is a particularly restrictive assumption to make in almost any applied work in the social sciences. As far as applications to the management of global climate problems goes, the restrictiveness of the assumption is particularly glaring. Uncertainty as to almost every thinkable aspect of the climate problem is highlighted in a qualified majority of applied work. In the history of game theory, moreover, we find many examples showing us that results may change in profound and interesting ways if we change our assumptions about the information structure of a particular game. In what follows one such a relaxation is discussed. It seems particularly relevant to the climate problem.

3.1 The Green-Porter model

The Green-Porter model takes as its departure the open horizon, *N*-person prisoner's dilemma game. What is new is that information is assumed to be 'publicly imperfect'. This is taken to mean that no player can observe the actions of any other players directly. However, each player can observe the same random variable, and this variable is related both to the actions taken in the game and to a random process unrelated to the actions of the players.

Some notation might be useful to clarify the idea. Let 'the quality of the atmospheric environment' at the end of period *t* be measured by an index given by the following expression $Q_t = \theta_t \sum_{i=1}^{N} c_{i,t}$. We take θ_t to be a random variable, which is assumed to be independently and identically distributed over time. This random variable captures 'natural fluctuations' in 'the quality of the environment.' The term $c_{i,t}$ is taken to be the abatement cost incurred by state *i* in period *t*. This cost is (for simplicity) assumed to be linearly related to improvements in 'the quality of the atmospheric environment', and the relationship may (without loss of generality) be assumed to have a unit slope. The incurred cost is a choice variable, and captures 'man-made fluctuations' in 'the quality of the environment'. What a player can observe is his own incurred abatement cost and the quality index. No player can observe the abatement cost incurred by any other player, or the realization of the random variable. Some further intuition is given by the following numerical example.

Let the range of θ_t be 0 to 1 inclusive. Furthermore, let the climate agreement contain 101 homogenous signatories, each of which has undertaken an obligation to make abatement efforts worth one cost unit each period. Non-compliance by signatory *i* in period *t* is taken to mean that *no* abatement costs was incurred by player *i* in period *t* ($c_{i,t} = 0$).

Suppose that the realized value of the random variable in period *t* was 1, and 50 states complied with the terms of the agreement in that period, the quality index becomes 50. The quality index also becomes 50 if the random variable realizes the value .5 and 100 states complied, or if all 101 states complied and the random variable realized the value 50/101, and so on. Thus, observing only the quality index and his own incurred abatement cost does not allow a player to draw firm conclusions about compliance. The reason is the intervention of "natural fluctuations" that create noise.

What then are the terms one can expect the players to comply with under such conditions? Green and Porter have shown that there exist equilibria in so-called trigger-level strategies. In a trigger-level strategy a player *i* complies as long as the random variable Q_t is kept at or above a constant threshold-value $\overline{Q}$. If the random variable falls below that threshold-value in a period *t*, player *i* ceases to comply in the *k* consecutive periods following *t*, independent

of the value that the random variable might take in these *k* periods. In period *t + k +1,* player *i* complies again. He thereafter complies in every period until the random variable falls below the threshold value again. From the period following immediately thereafter, he then repeats the same non-compliance pattern (with identical *k*).

In equilibrium the full group of players (*N*) play identical trigger-level strategies (same $\overline{Q}$, same *k*). A challenge is that there may (and most often will) exist more than a single pair of 'punishment phases' and threshold-levels ($\overline{Q}$ and *k*) that constitutes an equilibrium. Thus, we encounter a selection problem. A reasonable solution to this problem is to assume that players will coordinate on an agreement with a pair *k* and $\overline{Q}$ such that individual payoffs are maximized in equilibrium. Since players are treated symmetrically in equilibrium there is no distributive conflict, and this pair also maximizes aggregate payoffs under an equilibrium constraint. For this reason the selection criteria constitute a kind of collective rationality constraint, albeit very different from renegotiation-proofness.

A fascinating aspect of the equilibrium in trigger-level strategies is that no player ever cheats (this follows directly from the formulation of the equilibrium strategies). Periods of non-compliance are always triggered by natural fluctuations in equilibrium. If 'punishments' are removed (*k*=0) altogether, every player will have an incentive to cheat, and trigger-level equilibria with abatement levels higher than the dominant abatement level of the one-shot game do not exist.

What can be said about the trade off between the length of the punishment phase and the ambition of the threshold level? If we let the individual abatement efforts - $c_{i,t}$ - vary (which is possible within the Green-Porter framework) the answer is intuitive. Decreasing $\overline{Q}$ and/or decreasing *k* allows the abatement level to increase in equilibrium. This is so since both kinds of changes decrease the 'expected punishments'. Decreasing $\overline{Q}$ makes it less likely that the threshold falls short of actual quality, thus reducing the frequency of punishment phases. This is illustrated in figure 2 for a uniform distribution of θ_t on the interval [0,1].

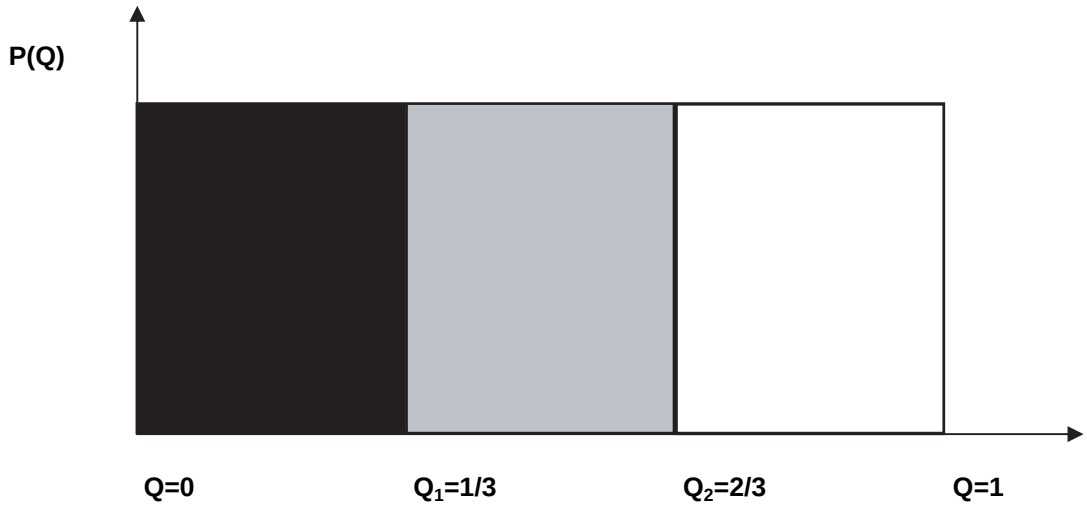

Figure 2. Probability of the environment having a particular quality

In figure 2, the probability of realizing an environmental quality at least as good as the threshold level of $\overline{Q_1}$ (in any given period) is given by the area of the distribution to the right of Q_1. This probability can be seen to be 1-(1/3)=2/3. Setting a more ambitious threshold-level at $\overline{Q_2}$, the probability of realizing a quality as least as good as the threshold-level corresponds to the area right of Q_2. This probability is 1-(2/3)=1/3. And in general, the more ambitious the threshold-level is, the smaller the probability that it will be achieved.

Decreasing *k* reduces the length of a punishment phase. In the efficient equilibrium (the one that maximizes individual payoffs) the pair ($\overline{Q}$, *k*) is selected in order to balance the cost of 'expected punishments' with the need to deter non-compliance.

For any equilibrium level of abatement efforts $c^*_{j,t}$ ($j \neq i$), it is possible for player *i* to calculate the probability that the actual quality falls short of the threshold-value if he chooses a non-equilibrium abatement level $c_{i,t}$. This is so since the distribution of θ_t (but not its realizations) is common knowledge. The probability of a specific realization of the quality is then

$$P_t(Q) = \theta_t\left((N-1)c^*_{j,t} + c_{i,t}\right)$$

The expression is clearly increasing in *N*: the number of players that the public good is defined upon, or less stringent, the number players that (for exogenous reasons) are *expected* to contribute to the provision of the good. So, for a given threshold-value, the probability that non-compliance will trigger a 'punishment phase' decreases with the number of players the good is defined upon (*N*). This is so since the marginal impact of a deviation on $P_t(Q)$ decreases with *N*, and thereby weakens the incentive to comply.

In a large group therefore, the efficient agreement must be less ambitious as far as environmental quality goes, or must stipulate longer phases of non-compliance after failing to attain the threshold value of environmental quality. It may be interesting to note in passing that a large group of countries were defined out of the hard-core obligations stipulated in the early in the Kyoto process (the Annex B countries).

Another interesting insight can be drawn from the expression above. Reducing the spread of the distribution θ_t will decrease the probability that a realization of a specific environmental quality falls below an ambitious threshold value because of natural fluctuations. Or in other words, reducing the spread of θ_t is the model expression of improved verification techniques, and such improvement allows for more ambitious agreements in equilibrium.

3.2 Relevance

Even though the Green-Porter model shares many restrictive assumptions with Barrett's model - notably assumptions (i), (iii), (iv), (vi), (vii) and (viii) in section 2.2 - it nonetheless relaxes the assumptions of complete information, instantly observable abatement levels and binary choice of abatement levels. So, in terms of realism one is inclined to say that a lot has been gained.

Note that the Green-Porter model has a size effect that relates to the one in Barrett's model. There are important differences, however. In Barrett's 1999 model the size effect is unrelated

to the *potential* group of signatories (N).[11] Low abatement levels are due to the restrictions placed on the number of *actual* signatories to an agreement by an ambiguous notion of collective rationality, which in turn is determined by the cost and benefit functions. This notion leads to counterintuitive conclusions. In the Green-Porter model the size effect relates to the size of the group of signatories, which is determined exogenously to the model. The last of these facets should, of course, be seen as a weakness.

Barrett's model determines only the umber of signatories to a self-enforcing climate agreement. The model is silent as far as the identity of the signatories goes. This is a weakness since it is profitable to not become a signatory, and since the selection problem must find its solution in any real world situation, be it in the laboratory or elsewhere. In the Green-Porter model there is no similar selection problem to be solved, although there is an equilibrium selection problem. This selection problem is, however, solved in a fairly uncontroversial way, by using a correspondence between individual and collective rationality.

3.3 Discussion

First and foremost the Green-Porter model seems attractive in that it produces predictions that - at least for me - are more intuitive. This is especially so on three points. In an environment with imperfect public information, cooperative efforts to provide collective goods may fail even though no party intentionally cheats. This is due to suspicion. Lowering the 'punishments' for failure to reach an environmental goal may extract higher abatement efforts in periods where the players abate, but will lead to less ambitious environmental goals and/or longer periods where emissions are uncurbed. This is due to a balance of expected cost and benefits. Finally, improved verification extracts more abatement.

This last aspect of the model is especially interesting with respect to the model for monitoring concentrations of greenhouse gases described in chapter 7 of the book. 'Environmental quality' is far from being a clear cut concept. If by natural fluctuations we for instance mean 'variations in mid temperature on the earth', the Green-Porter model seems uninteresting because of assumption (v) in section 2.2: The time lag from emissions to changes in mid temperature is very long and uncertain. Linking 'environmental quality' directly to emissions seems much more promising. Basing verification solely on information provided by the signatories, however, is risky because of incentives and opportunities to misrepresent. A monitoring system like the one outlined in chapter 7 of the book avoids both the lags and the misrepresentation problems, and fits nicely into what insights there are to be drawn from the Green-Porter model. This monitoring system attempts to allow verification by monitoring the concentrations of greenhouse gases in the atmosphere.

Again the most interesting evaluation of the model would likely be controlled laboratory tests of its central implications. There may even be some room for running tests of competing hypotheses in the Green-Porter and the Barrett model. I say more about this in the next section.

4 Strategies for empirical evaluation

As noted, no field data exist that can be used to evaluate different predictions about compliance with and verification of obligations undertaken by states in a global climate agreement. This is so simply because the agreement has not yet entered into force.

Lack of field data, however, does not preclude any kind of meaningful confrontation of game theoretic predictions in this field. It is possible to generate data in carefully designed

[11] In Barrett's 2002 paper, a link between N and equilibrium behaviour is identified.

laboratory settings. Experimental methods - like any method - have their strengths as well as their weaknesses (see the introduction to Kinder & Palfrey 1993 for good discussion).[12] Considering the strengths, even in situations where field data exist, it may well be worth the effort to 'triangulate' methodologically by supplementing analyses of field data with analyses of experimentally generated data.

In what follows a number of testable hypotheses derived from Barrett's model, as well as from the Green-Porter model, are sketched. They are partly overlapping, and partly competing. In my judgement they all lend themselves to experimental evaluation, but I have little to say here about the detailed design of such evaluations. That remains a task for the future.

In an environment of the kind described by Barrett, one should expect that:

H1: Increasing the cost (c) of cooperation increases cooperation (increases the number of subjects willing to enter into the agreement and fulfil their obligations)

H2: Increasing the marginal gain of cooperation (f) reduces cooperation (reduces the number of subjects willing to enter into the agreement and fulfil their obligations)

H3: Increasing the marginal gain of defection (g) increases cooperation (increases the number of subjects willing to enter into the agreement and fulfil their obligations)

H4: All of the above effects are amplified by allowing cheap talk between rounds.[13]

H5: None of the above is dependent on the size of the potential group of signatories (the total number of subjects in the experiment) for given parameter values.

In an environment of the kind described by Green and Porter one should expect that:

H6: Subjects select threshold-values and lengths of punishment phases close to the efficient equilibrium when both parameters are free.

H7: Subjects increase (decrease) their threshold-values when the length of the punishment phase is reduced (increased) from the efficient equilibrium.

H8: Subjects increase (reduce) the length of the punishment phase when the threshold-value is reduced (increased) exogenously from the efficient equilibrium.

H9: Subjects cooperate in more (fewer) periods when the spread of the probability distribution over natural fluctuations is reduced (increased).

H10: Cooperation decreases (in the sense of being seen in fewer periods) when the total number of subjects in the experiment is increased.

[12] Effective control and increased possibility of isolating the often particularistic relationships postulated by game theory are among the main advantages. The ability to 'create' un-observables like preferences and beliefs are also among the strengths. Prominently figuring among the weaknesses is the nagging question relating to the generalizability of results from the stylized laboratory settings (with weak financial incentives and unique attention to the problem at hand) to settings outside of the laboratory.

[13] Barrett (1999) stresses that cheap talk (especially by diplomats) between rounds facilitates the kind of collective rationality constraint (weakly renegotiation proof equilibria, WRPE) that he uses to arrive at his model conclusions. In this Barrett is in disagreement with van Damme 1989:207, note 1 (who considers WRPE in two player games): '[explicit communication] is irrelevant: Even if no player can articulate the proposal, the logic underlying the argument should convince both players not to punish each other (and themselves)'. Thus it is not obvious that H4 follows from Barrett's model.

As I have already made clear, I find H1, H2, and H3 to be particularly counterintuitive. In contrast to this I find H6, H7 and H8 more in line with intuition. In my view it would be particularly interesting to evaluate H9, considering the scientific model outlined in chapter 8 of the book.

H5 cannot be said to be counterintuitive, but it is rather unconventional considered as a formulation of the size principle (cf. Olson 1965). H10, on the other hand, deals with the total group of individuals that the good in question is defined upon, and is in this sense closer to Olson. However, since Olson makes a static argument, less cooperation means lower contributions per capita in a particular round. Lower sum contributions per capita over several rounds seems like an acceptable generalization to a dynamic environment. Both hypotheses are, however, interesting formulations of a size principle. It is especially interesting to note that previous experiments have found that per capita contributions increase with increasing group size (total number of participating subjects), controlled for the marginal productivity of a contribution (see the fascinating design and results in Isaak et al. 1984). Since H5 and H10 are clearly competing, and since previous research seems to cast doubt on H10, experimental evaluation of these two hypotheses may potentially add to our understanding of group interactions in public goods provision.

5 Conclusions

We now return to our central question, *is* the number of parties to an agreement a case of the 'more the merrier'? Repeated game models arguably provide us with well founded and precise *potential* answers – or conjectures – to this question. The two models discussed here answer in the negative. In this they stand firmly in a long and solid tradition, running back (at least) to Mancur Olson (1965).

It is interesting to note that Olson's size principle (i.e., the probability of some amount of collective good being provided is larger in smaller groups) does not show up in carefully designed experimental tests (Ledyard 1995, Palfrey 1993, Dawes et al. 1993, Isaak et al 1993). We may legitimately wonder if alternative formulations of the size principle, like the ones encountered in Barrett's model or the Green Porter model, will fare better in controlled laboratory tests.

In my opinion it is desirable, and maybe even necessary, to correct our theory building efforts in this field by data confrontations. Until we have done so, it is hard to say whether more *is* merrier, or not. Since the question is a very important one, I end with a plea for model testing.

Literature

Barrett S (2002) Consensus Treaties. *Journal of Institutional and Theoretical Economics* (forthcoming).

Barrett S (1999) A Theory of Full International Cooperation. *Journal of Theoretical Politics* 11:519-41.

Barrett S (1994) Self enforcing international environmental agreements. *Oxford Economic Papers* 46:878-94.

Bauer A (1992) *International cooperation over greenhouse gas abatement*. Mimeo. Seminar für empirische Wirtschaftsforshcung. University of Munich.

Carraro C & D Siniscalco (1993) Strategies for the International Protection of the Environment. *Journal of Public Economics* 52:309-28.

Carraro C & D Siniscalco (1992) The International Dimension of Environmental Policy. *European Economic Review* 36:379-87.

Dawes R et al. (1993) Organizing Groups for Collective Action. In: T Palfrey & D Kinder (eds.) *Experimental Foundations of Political Science.* Ann Arbor: The University of Michigan Press.

Farell J & E Maskin (1989) Renegotiation in Repeated Games. *Games and Economic Behavior* 1:327-60.

Finus M (2001) *Game Theory and International Environmental Cooperation*. Cheltenham: Edward Elgar.

Finus M & B Rundshagen (1998) Toward a positive theory of coalition formation and endogenous choice in global pollution control. *Public Choice* 96:145-86.

Fudenberg D & J Tirole (1991) *Game Theory*. Cambridge Mass.: The MIT-Press.

Green E & R Porter (1984) Noncooperative Collusion under Imperfect Price Information. *Econometrica* 52:87-100.

Hoel M (1992) International Environment Conventions: The Case of Uniform Reductions of Emissions. *Environmental and Resource Economics* 2:141-59.

Isaak R *et al*. (1993) Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations. In D Kinder & T Palfrey (eds.) *Experimental Foundations of Political Science*. Ann Arbor: The University of Michigan Press.

Kinder D & D Palfrey (1993) *Experimental Foundations of Political Science*. Ann Arbor: The University of Michigan Press.

Kong-Pin C (1995) On Renegotiation-Proof Equilibrium under Imperfect Monitoring. *Journal of Economic Theory* 65:610-23.

Ledyard J (1995) Public Goods: A Survey of Experimental Research. In: J Kagel & A Roth (ed.) *The Handbook of Experimental Economics*. Princeton: Princeton University Press.

Olson M (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups.* Cambridge Mass.: Harvard University Press.

Palfrey T (1993) The Conflict between Private Incentives and the Common Good. In: T Palfrey & D Kinder (eds.) *Experimental Foundations of Political Science*. Ann Arbor: The University of Michigan Press.

Rasmusen E (1994) *Games and Information* (2nd ed.) Oxford: Blackwell.

Schelling T (1992) Some Economics of Global Warming. *American Economic Review* 82 [1]:1-14.

Torvanger A (2001) *An analysis of the Bonn agreement: Backgroud implications for evaluating business implications*. Oslo: CICERO Report 2001:3.

UNEP 2001. *Climate Change 2001. Synthesis Report. Contribution of Working Groups I, II and III to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Van Damme E (1989) Renegotiation-proof Equilibria in Repeated Prisoners' Dilemma. *Journal of Economic Theory* 47:206-17.

Wagner U (2001) The Design of Stable International Environmental Agreements: Economic Theory and Political Economy. *Journal of Economic Surveys* 15:377-411.