NBER WORKING PAPER SERIES

CAN NEGOTIATING A UNIFORM CARBON PRICE HELP TO INTERNALIZE THE GLOBAL WARMING EXTERNALITY?

Martin Weitzman

Working Paper 19644
http://www.nber.org/papers/w19644

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2013

For useful constructive comments, but without necessarily implicating them in the conclusions of this paper, I would like to thank Frank Ackerman, Joseph Aldy, Geir Asheim, Daniel Bodansky, Daniel Cole, Stephen DeCanio, Denny Ellerman, Bard Harstad, Gary Libecap, Gilbert Metcalf, Juan Moreno-Cruz, Ian Parry, Robert Pindyck, Thomas Schelling, Robert Stavins, and David Weisbach. This paper was supported by a grant from the National Science Foundation. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Can Negotiating a Uniform Carbon Price Help to Internalize the Global Warming Externality?
Martin Weitzman
NBER Working Paper No. 19644
November 2013
JEL No. Q5,Q54

ABSTRACT

Thus far, most approaches to resolving the global warming externality have been quantity based. With *n* different national entities, a meaningful comprehensive treaty involves negotiating *n* different binding emissions quotas (whether tradeable or not). In post-Kyoto practice this *n*-dimensional coordination problem has proven intractable and has essentially devolved into sporadic regional volunteerism. By contrast, on the price side there is a natural one-dimensional focus on negotiating a single binding carbon price, the proceeds from which are domestically retained. Significantly (and unlike negotiated quantities) the negotiated uniform price on carbon emissions embodies an automatic "countervailing force" against free-riding self interest by incentivizing agents to internalize the externality. The model of this paper indicates an exact sense in which each agent's extra cost from a higher emissions price is counter-balanced by that agent's extra benefit from inducing (via the higher emissions price) all other agents to simultaneously lower their emissions. With some further restrictions, the theoretical model shows that population-weighted majority rule for a uniform price on carbon emissions can come as close to global efficiency as the median marginal benefit (per capita) is close to the mean marginal benefit (per capita).

Martin Weitzman
Department of Economics
Harvard University
Littauer 313
Cambridge, MA 02138
and NBER
mweitzman@harvard.edu

1 Background and Introduction

Throughout this paper I use the terms "climate change" and "global warming" interchangeably. The term "climate change" is currently in vogue and is perhaps a more apt description overall. But the term "global warming" is more evocative of this paper's main theme. Global warming is a *global* public-goods externality whose resolution requires an unprecedented degree of international cooperation and coordination. This international climate-change externality has been characterized as the biggest public goods problem that humanity has ever faced. I concentrate in this paper on carbon dioxide emissions, but in principle the discussion could be extended to emissions of all relevant greenhouse gases.

An internationally harmonized but nationally retained carbon tax has already been proposed as a solution to the global warming externality, and has been examined on its merits.[1] In what follows I very briefly summarize some of the possible virtues of an internationally-harmonized nationally-collected carbon tax that have been noted in the literature. My foil here is an internationally harmonized cap-and-trade system. This kind of global-design comparison is complicated and full of subjective judgements about what might or might not work better in practice and why or why not. My purpose here is merely to indicate that a carbon tax already has some significant arguments in its favor – as a prelude to some new theoretical arguments for negotiating a uniform price on carbon that I will later develop in this paper.

Both quantity-based and price-based controls are inherently uncertain for the period during which they apply (in between times of periodic review), but the uncertainty takes different forms. With cap-and-trade, total emissions are known but the price is uncertain. With a carbon tax, the price of carbon emissions is known, but total emissions are uncertain. On the basis of economic models of climate change that include uncertainty, carbon taxes outperform tradeable permits, both empirically and theoretically.[2] In the real world, I think that energy price volatility is very poorly tolerated by the general public. Swings in carbon prices, especially in extreme cases, could sour the public and discredit for some time the entire idea of a market-based approach to the climate change problem. On the other hand, it is difficult for me to imagine the broad public getting quite so upset because total emissions are uncertain.

It has been argued, I think convincingly, that a carbon tax is more easily administered and more transparent than a cap-and-trade system. This consideration is especially important in a comprehensive international context that would include all major emitting countries.

[1] There is actually a fair-sized literature on a carbon-tax approach. See, e.g., Cooper (2010), Metcalf and Weisbach (2009), Nordhaus (2007, 2013), and the many further references cited therein.

[2] See Hoel and Karp (2002), Pizer (1999), and Weitzman (1974).

Under international cap-and-trade, governments will allocate valuable emissions permits to their nation's firms and residents. The incentive for kleptocrats to steal these valuable emissions permits and sell them on the international market is presumably much more rewarding than the temptation to not enforce an internal tax on emissions.

The revenues from an internationally harmonized carbon tax are retained internally within each nation, and could be used, for example, to offset other taxes. This, I think, is a desirable property. By contrast the revenues generated from an internationally harmonized cap-and-trade system flow as visible external transfer payments across national borders, which might be less easily tolerated by nations needing to pay other nations large sums of taxpayer-financed money to buy permits.

This extremely brief discussion of the advantages of an internationally harmonized carbon tax (compared to cap-and-trade) was never intended to be comprehensive. There are also some legitimate arguments in favor of internationally harmonized tradeable permits and against a carbon tax.[3] Both approaches are subject to immense – sometimes seemingly overwhelming – criticisms. In both cases there are innumerable practical details that must be worked out. In both cases an effective international treaty needs to be binding, which raises uncomfortable issues of enforcement mechanisms and international sanctions. Additionally, there might be mixed hybrid systems. I merely wanted to establish a starting position where an internationally harmonized carbon tax already commands some intellectual respect.

The Kyoto approach to global warming was inspired by the ultimate vision of a top-down worldwide treaty limiting the output of each nation's carbon dioxide emissions. It had been wishfully hoped that the highly incomplete Kyoto quantity assignments might have grown into a comprehensive binding system of national emissions caps. If these comprehensive caps were freely traded internationally as emissions permits, it would have caused there to be one uniform worldwide price of carbon emissions, thereby guaranteeing cost effectiveness.

As events played out, Kyoto did not come close to its inspirational vision of an internationally harmonized binding system of emissions caps. By now, the quantity-based Kyoto-type approach has pretty much broken down, leaving the world with a highly non-optimal patchwork of sporadic regional volunteerism that does not address centrally how to correct the critical externality of global warming.

The primary lesson I take away from the breakdown of Kyoto is the need for fresh thinking, new insights, and, perhaps, a different global-design approach to the externality problem. In this paper I examine the theoretical properties of a natural one-dimensional focus on negotiating a single binding price on carbon emissions, the proceeds from which are

[3] For a critical review of carbon taxes vs. cap-and-trade, see Goulder and Schein (2013) and the further references they cite.

domestically retained. For simplicity, I identify this single binding price on carbon as if it is a harmonized carbon tax. At a theoretical level of abstraction, I blur the distinction between a carbon price and a carbon tax. However, in actuality the important thing is acquiescence by each nation to a binding minimum price on carbon emissions, not the particular mechanism by which this binding minimum price is attained by a particular nation. I elaborate further on this issue in my concluding remarks.

At a theoretical level, I would suggest that the instruments of negotiation for helping to resolve the global warming externality should ideally possess three desirable properties.

1. *Induce cost effectiveness.*

2. Be of *one dimension centered on a "natural" focal point* to facilitate finding an agreement with relatively low transactions costs.

3. *Embody "countervailing force" against narrow self interest by automatically incentivizing all negotiating parties to internalize the externality.*

Using these three desirable theoretical properties as criteria, I now compare and contrast an idealized binding harmonized price with an idealized binding cap-and-trade system.

On the first desirable property, in principle both a carbon price and tradeable permits achieve cost effectiveness (provided agreement can be had in the first place).

The second desirable property (low dimensionality) argues in favor of a one-dimensional harmonized carbon price over a n-dimensional harmonized cap-and-trade system. Alas, this argument is elusively difficult to formulate rigorously, or even to articulate coherently. My argument here is necessarily intuitive or behavioral and relies on empirical counter-examples. In this case a primary empirical counter-example is the breakdown of the quantity-based Kyoto approach.

With n different national entities, a quantity-based treaty involves assigning n different binding emissions quotas (whether tradeable or not). Quantity-based treaty making can be viewed as a coordination game with n different players. Such a game can have multiple solutions, often depending delicately on the setup and what is being assumed. In the case of Kyoto, the world has in practice arrived at a bad solution that has essentially devolved to regional volunteerism.

Thomas Schelling introduced and popularized the notion of a focal point in game theory.[4] Generally speaking, a focal point of a n-party coordination game is some salient feature that reduces the dimensionality of the problem and simplifies the negotiations by limiting bargaining to some manageable subset, hopefully of one dimension. The basic idea is

[4] Schelling (1960). See also the special 2006 issue of the *Journal of Economic Psychology* devoted to Schelling's psychological decision theory, especially the introduction by Colman (2006). Three of the seven articles in this issue concerned aspects of focal points, testifying to the lasting influence of the concept.

that by limiting bargaining to a salient focus, there may be more hope of reaching a good solution. In a somewhat circular definition, a focal point is anything that provides a focus of convergence. The "naturalness" or "salience" of a focal point is an important aspect of Schelling's argument that is difficult to define rigorously and is ultimately intuitive.

The concept of "transactions cost" is associated with the work of Ronald Coase.[5] The basic idea is that n parties to a negotiation can be prevented from attaining a socially desirable outcome by the costs of transacting the agreement among themselves. One could try to argue that, other things being equal, transactions costs increase at least proportionally with the number of parties n.

In the case of international negotiations on climate change, I believe that both Schelling's concept of a salient focal point and Coase's concept of transactions costs can be used as informal arguments to support negotiating a single harmonized carbon price whose proceeds are nationally rebated. Put directly, it is easier to negotiate one price than n quantities – especially when the one price can be interpreted as "fair" in terms of equality of effort. I cannot defend this claim rigorously. At the end of the day, this is more of a plausible conjecture than a rigorous theorem. Whether justly or not, throughout this paper I basically assume that the essential contrast is between one binding price assignment versus n binding quantity assignments – and I then proceed to examine the consequences.[6]

The third desirable property is that the instrument or instruments of negotiation should embody "countervailing force" against narrow free-riding self interest by incorporating incentives that automatically internalize the externality. I believe this third property is arguably the most important property of all. This "countervailing force" property is inherently built into a price-based harmonized national carbon price, but it is absent from a quantity-based international cap-and-trade system, at least as traditionally formulated.

If I am assigned a cap on emissions, then it is in my own narrow free-riding self interest to want my cap to be as large as possible (whether or not my cap will be tradeable as a permit). The self-interested part of me wants maximal leniency for myself. There is no countervailing force on the other side encouraging me to lower my desired emissions cap because of the externality benefits I will be bestowing on others.

Within a nation, the government *assigns* binding caps. But *among* sovereign nations,

[5] Coase himself did not invent or even use the term "transactions cost" but he prominently employed the concept. See Coase (1960). For an application of the transactions cost approach to controlling greenhouse gas emissions, see Libecap (2013).

[6] Later I discuss conceivable attempts to reduce the dimensionality of negotiating n quantities to negotiating one aggregate quantity (which would involve *two* rounds of negotiations). In the end I conclude that such attempts will likely founder on the same underlying Kyoto-like problem of negotiating the n underlying quantity-like entities in the first round that are required to construct the one-dimensional aggregate that is negotiated in the second round.

binding caps must be *negotiated*. I believe that this is a crucial distinction for the success or failure of a cap-and-trade regime. A Kyoto-type quantity-based international system fails because no one has an incentive to internalize the externality and everyone has the self-interested incentive to free ride. What remains is essentially an erratic pattern of benevolent individual volunteerism that is far from a socially optimal resolution of the problem.

A internationally-harmonized domestically-collected carbon price is different. If the price were imposed on me alone, I would wish it to be as low as possible so as to limit my abatement costs. But when the price is uniformly imposed, it embodies a countervailing force that internalizes the externality for me. Countervailing my desire for the price to be low (in order to limit my abatement costs) is my desire for the price to be high so that other nations will restrict their abatements, thereby increasing my benefit from worldwide total carbon abatement. A binding uniform price of carbon emissions has a built-in self-enforcing mechanism that countervails free riding.[7]

The remainder of the paper concentrates mostly on analyzing this third "countervailing force" property of an internationally-harmonized but nationally-collected carbon price. I construct a basic model indicating the exact sense in which each agent's extra cost from a higher emissions price is counter-balanced by that agent's extra benefit from inducing all other agents to simultaneously lower their emissions.

With further restrictions, the model shows that population-weighted majority rule for an internationally harmonized carbon price can come as close to an optimal price on emissions as the median per-capita marginal benefit is close to the mean per-capita marginal benefit. The key insight from this way of looking at things is that in voting (or more generally negotiating) a universal carbon price, various nations are, to a greater or lesser degree, internalizing the externality. Loosely speaking, an "average" nation is fully internalizing the externality because its extra cost from a higher emissions price is exactly offset by its extra benefit from inducing all other nations to simultaneously lower their emissions.

On the price side, a uniform carbon price automatically has the desirable property that cost effectiveness is guaranteed. I think that the formal voting result of the model of this paper might perhaps be interpreted somewhat less formally as indicating that negotiating an internationally harmonized (but nationally collected) carbon price may have an important desirable property on the quantity side as well. If the median marginal benefit (per capita)

[7] Later I discuss negotiating one worldwide aggregate emissions quota (*contingent* upon a previous-round assignment of n fractional targets, set, for example, by a preceding agreement on various target reductions from various baselines). A system based on negotiating aggregate emissions could, in principle, embody countervailing force against the global warming externality. But, again, I will conclude that negotiating the extra layer of n first-round Kyoto-like fractional target reductions will likely founder politically when applied on a worldwide scale.

equals the mean marginal benefit (per capita), then the socially optimal carbon price has the property that, roughly speaking, half of the world's population wants the price to be higher, while the other half of the world's population wants the price to be lower. In this situation, the desirable quantity-side property is that the total worldwide output of all emissions might be "close" to being optimal to the extent that the outcome of negotiations mimics the outcome of majority voting. Although the real world is a far more complicated and nuanced place than the restrictive theoretical model of this paper, I think this voting result is trying to indicate something positive (even if only at an abstract level) about how a negotiated uniform carbon price might possess some overall potential to counteract via internalization the externality of global warming.

2 The Model

The formulation here is at a heroic level of abstraction. I wave away innumerable "practical" considerations to focus on a theoretical model. I beg the reader's indulgence for a willing suspension of disbelief while the basic argument is being developed.

The analysis is made cleanest and most transparent when the fundamental unit is the person, so that everything is normalized per capita. In reality, of course, people belong to some larger entity, here called a "nation," that (hopefully or presumably) acts on their behalf with respect to carbon price negotiations, enforcement, and revenue recycling. The nation here is an elastic concept, since for the purposes of this paper it might be more appropriate to consider regional blocs like the European Union as if it comprised a single nation. It is easiest to conceptualize that all of the people belonging to one nation are identical agents whose tastes and technology are representative of that nation. For an individual belonging to a nation everything – emissions, costs, benefits – is expressed in per-capita terms for that nation. (Inversely, one could take costs and benefits on the national level as given primitives and impute to each citizen the corresponding per-capita costs and benefits as a function of per-capita emissions, being careful to ensure that the imputed per-capita costs and benefits aggregate consistently to the given national costs and benefits.)

The nation here is effectively an entity that enforces the imposition of an internationally harmonized carbon price and recycles internally the domestic revenues raised by the price. I assume that this recycling is efficient, as if by lump sum internal transfers, so there is no net loss from the carbon price per se. Additionally, when it comes to voting or negotiating a carbon price for some particular time period, the nation effectively votes or negotiates on behalf of its citizens in accordance with their preferences. These assumptions are vulnerable, but they may make sense as an abstraction and can serve as a point of departure for further discussion.

The total world population is m. Each person is indexed by $i = 1, 2, ..., m$. In what follows I abstract away from dynamics in favor of a static-flow analysis. I assume agents can convert their wishes about desired stock levels into wishes about corresponding flows for the period under consideration.

Let X_i stand for the level of carbon abatement of person i (from some predetermined level). The cost of attaining abatement level X_i for person i is given by the function $C_i(X_i)$. If the internationally harmonized price on carbon emissions is p, then the response of individual i is $X_i(p)$, where, for each $i = 1, 2, ...m$,

$$C_i'(X_i(p)) = p. \tag{1}$$

Condition (1) guarantees worldwide cost effectiveness. The total worldwide abatement level corresponding to (1) is

$$X(p) = \sum_{i=1}^{m} X_i(p). \tag{2}$$

The benefit of worldwide abatement level X for person i is given by the benefit function $B_i(X)$. The worldwide socially optimal level of an internationally harmonized emissions price is the value p^* that obeys the classic Samuelson public goods optimality condition, which here can be written as

$$p^* = \sum_{i=1}^{m} B_i'(X(p^*)). \tag{3}$$

Consider next what is the optimal level of an internationally harmonized carbon price from the narrow perspective of person i. Because revenues from the carbon price are collected and recycled by the nation to which i belongs, there is no tax burden per se. (The only real burden to i here is the cost C_i incurred by obeying condition (1)). The emissions-price level p_i that i would most prefer solves the problem

$$\max_p \{B_i(X(p)) - C_i(X_i(p))\}, \tag{4}$$

which satisfies the first-order condition

$$B_i'(X(p_i))\, X'(p_i) = C_i'(X_i(p_i))\, X_i'(p_i). \tag{5}$$

Use condition (1) to rewrite (5) as

$$p_i = C_i'(X_i(p_i)) = \lambda_i\, B_i'(X(p_i)), \tag{6}$$

where

$$\lambda_i \equiv \frac{X'(p_i)}{X_i'(p_i)} = \frac{dX}{dX_i} \tag{7}$$

might be called the *externality-internalizing multiplier* (for agent i).

Note from (6) what agent i is *not* doing here. Agent i is *not* equating its marginal cost of abatement C_i' to the narrow marginal benefit from one more unit of its own abatement B_i', which would be the analogue here to the condition for voluntary provision of public goods, and which would result in a free-riding too-low provision of the public good. Instead, the narrow marginal benefit B_i' is being magnified in (6) by a factor of λ_i, so that agent i is equating its marginal cost C_i' to $\lambda_i\, B_i'$ (instead of to B_i').

What is the value of the externality-internalizing multiplier λ_i? If all agents i were identical, then $\lambda_i = m$ for all i and the scaling-up multiplier per-capita is world population. In the more general case, by (7) the multiplier λ_i is the ratio of the change in total global marginal abatement dX divided by the change in agent i's marginal abatement dX_i. Equation (6) (along with definition (7)) signifies that agent i is internalizing the externality that it is causing by applying a multiplier that scales up the effect of its narrow marginal benefit by however many times greater is the world's marginal abatement response (to a price change) than i's own marginal abatement response (to a price change). Effectively, agent i is induced to scale up its own narrow marginal benefit to a kind of golden-rule-like imputation of the corresponding worldwide marginal benefit. But this is just the kind of golden-rule-like scaling-up property that we would want an externality-internalizing multiplier to possess.

While it is simple, equation (6) (along with definition (7)) is a fundamental result of this paper. It conveys the exact sense in which a uniform national carbon price is internalizing the global warming externality. Again, the basic idea is that each agent's extra cost from a higher uniform emissions price is counter-balanced by that same agent's extra benefit from inducing all other agents to simultaneously lower their emissions. This critical counter-balancing incentive is transmitted via the externality-internalizing multiplier.

If all agents have identical cost and benefit functions, then $\lambda = m$, while $X_i = X/m$ and $p_i = p$, so that (6) becomes

$$p = C_i'(X_i) = m\, B_i'(X) \tag{8}$$

for all i, which is exactly the classic Samuelson condition for public goods optimality with m identical agents.

One might try, heuristically, to make a more general statement than (8) about an "average" agent along similar lines. Loosely speaking, if i^* is an "average" citizen of the world (in a sense yet to be defined), one might be tempted to allow the approximation λ_{i^*} $(=X'(p_{i^*})/X_{i^*}'(p_{i^*})) \approx m$. Speaking even more loosely, one might be further tempted to envision, for this "average" citizen of the world i^*, that $B'_{i^*}(X(p_{i^*}))$ is an "average" value of all $\{B'_i(X(p_{i^*}))\}$. Having come this far, the ultimate temptation is to reason super-loosely that p_{i^*} from (6) then might not be a terrible approximation for p^* from (3). Such an argument is heuristic and crude, to put it mildly. To make this kind of an argument about an "average" agent more precise requires placing considerably more structure on the cost and benefit functions.

3 Some Further Simplifying Assumptions

The formulation in the last section gives some useful broad insights about the externality-internalizing multiplier, but expression (6) (along with definition (7)) is too general to yield tractable analytical solutions. I proceed to get sharper closed-form expressions by considering families of linear marginal cost functions and linear marginal benefit functions, all members of which are restricted to having identical slopes, but each member of which can have a different intercept representing differing values of an individual shift parameter. This is the simplest formulation that allows costs and benefits to be different yet delivers analytically tractable results.

Without further apologizing, I assume for all persons $i = 1, 2, ..., m$ that marginal costs are restricted to be of the particular linear form

$$C'_i(X_i) = c_i + \gamma X_i. \tag{9}$$

The simplification (9) corresponds to a family of linear supply schedules having the same slope γ that are shifted up or down (or right or left) to various degrees for various different individuals. Condition (9) means that marginal costs are linearly symmetric in such a way that the coefficient c_i gives an unambiguous ranking of marginal costs for any arbitrarily-given common level of abatement. Everyone has an individually shifted version of the same underlying linear schedule of marginal cost (or linear supply schedule). Henceforth we can identify the marginal cost schedule of person i as being represented by c_i (given the common value of γ).

Also without further apologies, it is assumed for all persons i that marginal benefits are restricted to be of the particular linear form

$$B'_i(X) = b_i - \beta X. \tag{10}$$

Here the simplification (10) corresponds to a family of linear demand curves having the same slope -β that are shifted up or down (or right or left) to various degrees for various different individuals. Condition (10) means that marginal benefits are linearly symmetric in such a way that the coefficient b_i gives an unambiguous ranking of marginal benefits for any arbitrarily-given level of total abatement. Everyone has an individually shifted version of the same underlying linear schedule of marginal benefit (or linear demand schedule). In this sense we can henceforth identify the marginal benefit schedule of person i as being represented by b_i (given the common value of β).

Without specifications amounting to shifted linear supply schedules and shifted linear demand schedules, it is very difficult to obtain neat results. I think that the formulation of this section may be all right as a base case or point of departure. The next section obtains some strong insights that can emerge from assuming (9) and (10).

4 A Majority-Rule Result

Plugging (9) into (1) yields

$$c_i + \gamma X_i(p) = p, \tag{11}$$

which can be inversely solved to obtain the relevant response function

$$X_i(p) = \frac{p - c_i}{\gamma}. \tag{12}$$

Combining (12) with (2) gives

$$X(p) = \frac{mp - \sum c_i}{\gamma}. \tag{13}$$

To obtain the socially optimal p^*, plug (13) and (10) into (3), which turns (3) into the equation

$$p^* = \sum b_i - m\beta \left(\frac{mp^* - \sum c_i}{\gamma} \right). \tag{14}$$

Finally, inversely solve equation (14) for p^*, which can then be expressed in the form

$$p^* = k\bar{b} + k', \tag{15}$$

where

$$\bar{b} \equiv \frac{\sum b_i}{m}, \tag{16}$$

while

$$k \equiv \frac{m\gamma}{\gamma + m^2\beta}, \tag{17}$$

and

$$k' \equiv \frac{m\beta \sum c_i}{\gamma + m^2 \beta}. \tag{18}$$

To obtain the individually optimal p_i, first note from (13) and (12) that

$$\frac{X'(p_i)}{X_i'(p_i)} = m. \tag{19}$$

Then substitute (19), (13), (10) into (6), which turns the latter expression into

$$p_i = \left[b_i - \beta \left(\frac{mp_i - \sum c_i}{\gamma} \right) \right] m. \tag{20}$$

Finally, inversely solve equation (20) for p_i, which can then be expressed in the form

$$p_i = kb_i + k', \tag{21}$$

where, as before, k is defined by (16) and k' is defined by (17).

Equation (21) means that the ordering of preferred carbon prices is the same as the ordering of marginal benefits. In this particular linear setup, it turns out that marginal costs $\{c_i\}$ are internalized and do not play a role in the *comparative* ranking of preferred carbon prices (because k is independent of costs), although they do play a role in the *absolute* level of preferred carbon prices (via their aggregate influence on k').

Note the tight correspondence between (21) and (15). To explore this correspondence further, imagine the following thought experiment.

Waving aside how it came into existence, suppose there is a World Climate Assembly (WCA). The WCA votes on pairwise alternatives for the desired level of a universal carbon price, based on the principle of one-person one-vote. In practice, this means that nations vote for their desired level of a universal carbon price on behalf of their citizen constituents, but the votes are weighted by each nation's population.

What is the justification for a new international organization like the WCA? The ultimate justification is that new big problems may require new big solutions. For a world desperately wanting new solutions to the important externality of climate change, perhaps it is at least worth considering establishing a new organization along the lines of WCA. After all, it is useful to have some concrete fallback decision mechanism behind vague "negotiations" because even with the focus on a one-dimensional harmonized carbon price there are bound to be disagreements, whose resolution is unclear. I merely assume that it is in the interest of enough nations to forfeit their rights to pollute in favor of a WCA solution of the global warming externality. Taken less literally, the thought experiment of a hypothetical WCA can still help us to concentrate our thinking and intuition on what negotiations should be trying to accomplish.

With pairwise majority voting on the preferred value of p, by the median voter theorem the equilibrium outcome will be the median value of $\{p_i\}$, here denoted $\widehat{p}$. Let $\widehat{b}$ denote the median value of $\{b_i\}$. Then by (21), the majority-preferred value of p is the value $\widehat{p}$ satisfying

$$\widehat{p} = k\widehat{b} + k'. \tag{22}$$

Compare (22) with (15). The majority-rule carbon price $\widehat{p}$ is close to the optimal carbon price p^* when the median marginal benefit $\widehat{b}$ is close to the mean marginal benefit $\bar{b}$. This is as good a result as one might hope for from a voting solution. The mean and the median are both measures of central tendency. At this level of abstraction I find it difficult to argue whether the mean marginal benefit of abatement per capita should be greater or less than the median marginal benefit of abatement per capita. If the two are equal, then majority voting obtains the optimal solution. If the two are unequal, the analysis provides a measure of how far away from optimality is majority rule. Of course this is just a model with quite restrictive assumptions, but in a post-Kyoto world of stalemated negotiations I find attractive the image of a WCA-style population-weighted median carbon price as being a useful point of departure that holds out some prospect of coming "close enough" to an optimal solution.

This is about as far as theory can take us. When the model is tightly structured with the specifications and assumptions applying to this section of the paper, the main result here indicates an exact sense in which majority rule for a harmonized national carbon price can come close to fully and completely internalizing the global warming externality. As was previously indicated, I think that the formal WCA voting result of the model of this section of the paper may perhaps be interpreted somewhat less formally as indicating that negotiating a uniform national carbon price may have a desirable property that favors supplying a near-optimal level of emissions. If the median marginal benefit (per capita) is close to the mean marginal benefit (per capita), then the socially optimal carbon price has the property that roughly half of the world's population wants the price to be higher, while roughly the other half of the world's population wants the price to be lower. This might be interpreted as a desirable feature even without the formal mechanism of majority-rule voting in the WCA.

5 Might a Modified Cap-and-Trade Work as Well?

In the introduction I listed three desirable features that instruments for negotiating climate change should ideally possess: (1) cost effectiveness; (2) a natural one-dimensional focal point; (3) a built-in self-enforcement mechanism that internalizes the externality. I then explained that a harmonized national carbon price possesses all three properties, whereas a harmonized n-dimensional cap-and-trade system at best (if it can be negotiated in the first place) possesses only the first property of cost effectiveness. With n different nations, there will be difficult bargaining over n different caps with no force countervailing each nation's selfish desire to be a free rider and secure for itself a large cap on emissions.

But maybe I am being unfair to tradeable permits. Suppose we imagine trying to convert the n-dimensional problem of allocating carbon emissions permits into some one-dimensional quantity analogue of a uniform price on carbon emissions. We might imagine a thought experiment where the cap-and-trade negotiators are sitting around a negotiating table and limiting themselves to simple linear formulas.

For illustrative simplicity, suppose the cap-and-trade negotiators must decide the total amount of emissions and the fractional allocation of emissions for each nation. If Y is the worldwide total output of emissions, then country j might be assigned the *fraction* a^j of worldwide emissions according to the linear formula

$$Y^j = a^j\, Y, \tag{23}$$

where Y^j is the emissions cap assigned to country j, while a^j is some given distribution coefficient representing j's assignment fraction, with the property

$$\sum_{j=1}^{n} a^j \; = 1. \tag{24}$$

Given the assigned distributional fractions $\{a^j\}$ and the formula (23), one might then imagine negotiating over (or even voting for) the total emissions Y. This system seemingly possesses the desirable property of having a one-dimensional locus of negotiations (here Y). And there is also countervailing force against negotiating for a high value of Y. Although j's automatic assignment of a high emissions target Y^j when Y is high (via (23)) helps j directly, this domestic effect is counteracted by the benefits that j would lose from high Y because everyone else would then also emit more. It appears that such a cap-and-trade system could in principle have desirable focal-point and countervailing-force properties *if* the assigned fractions $\{a^j\}$ were accepted and bargaining were restricted to negotiating Y.

But now follow the thought experiment further by asking : Where do the coefficients $\{a^j\}$ come from in the first place? They are presumably the result of a n-dimensional negotiating process where there is no countervailing force to the selfish desire of country j to make its own a^j as high as possible. With n different nations, there will be the usual difficult bargaining over n different distributional fractions $\{a^j\}$, with no externality-internalizing incentive countervailing each nation's desire to secure for itself a high fraction of emissions – again presumably resulting in a Kyoto-like breakdown.

When a cap-and-trade system is used to control pollution *within* a nation, the government of that nation *assigns* the caps analogous to Y^j or the fractions analogous to a^j. (Often these quantities have been allocated for free based on some proportionality formula like a uniform reduction of previous pollution levels, which eases acceptance by the polluters.) In this intra-national case there is a natural symmetry between a one-dimensional price p and a one-dimensional total quantity Y. But there is no international government that has the unilateral power to assign caps or fractions. These caps or fractions must be *negotiated* among sovereign nations. This breaks the one-dimensional symmetry because now one price p is contrasted with the asymmetry of n vested sovereign interests jockeying for the n initial distributions of the form $\{Y^j\}$ or $\{a^j\}$. There is thus a critical distinction between intra-national and inter-national cap-and-trade systems. In the inter-national case the initial distribution of caps is explicitly distributive, resulting in a war of words about who caused the global-warming problem and who should bear the burden of remedying it, who is rich and who is poor, and so forth and so on.

But perhaps even this formulation is being unfair to cap-and-trade. We might try to imbue the $\{a^j\}$ with focal-like salient qualities by imagining "naturally symmetric" allocations of $\{a^j\}$. One such seemingly symmetric formula might be that each country is assigned the same fractional reduction of emissions from some agreed-upon baseline year. The Kyoto Protocol of 1997 adopted just a little of the spirit of this idea for developed countries alone, with the hope that some variant of it might later be extended to developing countries. The high-income industrialized countries (Annex I) agreed to "binding" commitments (but without any enforcement mechanism!) to reduce greenhouse gas emissions in 2012 by an average of 5% relative to 1990 levels (although allowing individually-negotiated variations around that 5% average). Developing countries were exempt from any "binding" commitments. Overall, the Kyoto Protocol did not come close to fulfilling its initial aspirations. The U.S. did not ratify, Canada dropped out, and individual compliance was at best spotty.[8]

[8] The one bright spot might be considered the European Union, whose emissions trading system could perhaps be interpreted as evolving towards an EU-wide cap (declining annually) with member-state shares increasingly being determined by auctioning permits. I am unsure and somewhat skeptical about the extent to which this EU model might be extended to the world as a whole. For a generally favorable assessment

Furthermore, and perhaps most distressingly, non-Annex I countries have not agreed to any future "binding" commitments going forward from 2012. The Kyoto experience is subject to multiple interpretations. For me, it largely testifies to the great difficulty of negotiating binding international caps on the major emitters. In the language of (23), it has been overwhelmingly problematic to assign binding quantity-like distributional coefficients $\{a^j\}$ on a worldwide basis.

Other seemingly symmetric quantity formulas might also be examined. For example, one might entertain the idea of assigning the same worldwide emissions level per capita. This is a symmetric formula that embodies a certain concept of worldwide fairness, but a cap-and-trade system based on such an initial distribution of caps would involve massive transfers from the developed to the developing countries that would likely prove politically unacceptable. Besides, even this formula does not address concerns regarding historical responsibility for the cumulative stock of emissions that would surely be raised. Alternatively, one might imagine negotiating (or even voting on) an identical worldwide percentage reduction from some base case of emissions. In this situation, I think, everyone would first argue about the baseline emissions that they were initially assigned.

I abstain from further speculation. My point is that no matter what quantity-like initial allocation mechanism I can imagine, an attempt to modify an international cap-and-trade system by making it one dimensional seems likely to founder for essentially the same reasons that an unmodified international cap-and-trade system founders. In a quantity-based system with n different sovereign nations I fear there will be intractable negotiations for n different distributional assignments, with no force countervailing each nation's free-riding desire to secure for itself a selfishly lenient emissions fraction.

Here is what I think is the essence of the one-price vs. n-quantities negotiation problem of this section. A quantity-type system based on a formula like (23) involves *two* layers of negotiations. First, the n parties must agree on the n quantity-like distributional coefficients $\{a^j\}$. Then, second, the parties must agree on the single aggregate level of Y. By contrast, a price-based system involves only *one* layer of negotiation, focused on agreeing to a single one-dimensional uniform price p. This latter is not an easy task, but it makes sense to me that it is generally easier to negotiate one price layer than two quantity layers (whose first layer involves assigning n quantity-like distributional coefficients $\{a^j\}$). Admittedly this argument depends upon a particular way of framing the issue, but it seems to me that, in international negotiations among n sovereign nations, there may be an irreducible asymmetry between one price instrument vs. n quantity instruments.

Even while acknowledging that it only involves one layer of negotiations (as opposed to

of this possibility, see Ellerman (2010).

two on the quantity side), one might ask on the price side what might induce n countries to agree on a single harmonized charge for carbon emissions. We have been over this ground before. There is no airtight logic, only a series of partial arguments. One argument is that the uniform price is nationally collected, so that the contentious distributional side is somewhat hidden and there is at least the appearance of fairness as measured by equality of effort. A second desirable feature, I have argued, is the natural salience and relatively low transaction costs of negotiating one price as against negotiating n quantities, which, while somewhat imprecise, is in my opinion an important distinction. A third argument is the self-enforcement mechanism that constitutes the main theme of this paper, namely the built-in countervailing force of an imposed uniform price of carbon, which tends to internalize the externality and gives national negotiators an incentive to offset their natural impulse to bargain for low tax rates for themselves.

Of necessity, this paper has been sprinkled with subjective judgements. This, unfortunately, is the nature of the subject. To repeat yet again, this time after examining somewhat more carefully the alternatives, I judge it difficult to escape the conclusion that, in the context of an international treaty that covers all major emitters, it is more politically acceptable and it comes closer to a social optimum to negotiate one binding price than n binding quantities or quantity-like distributional coefficients.

6 Concluding Remarks

The model of this paper is so abstract and so removed from reality that it is open to enormous amounts of criticism on many different levels. There are so many potential complaints that it would be incongruous to list them all and attempt to address them one by one. These many potential criticisms notwithstanding, I believe the model here is exposing a fundamental countervailing-force argument that deserves to be highlighted.

Because the model is at such a high level of theoretical abstraction, it has blurred the distinction between a carbon price and a carbon tax. As was previously noted, the important thing is acquiescence by each nation to a binding minimum price on carbon emissions, not the particular internal mechanism by which this obligation is met. A system of national carbon taxes with revenues kept in the taxing country is a relatively simple and transparent way to achieve harmonized carbon prices. But it is not necessary for the conclusions of this paper. Nations or regions could meet the obligation of a minimum price on carbon emissions by whatever internal mechanism they choose – a tax, a cap-and-trade system, a hybrid system, or whatever else results in an observable price of carbon.

The purpose of this paper is primarily theoretical. *Any* proposal to resolve the global warming externality will face a seemingly overwhelming array of practical administrative obstacles and will need to overcome powerful vested interests. That is the nature of the global warming externality problem. The theory of this paper seems to indicate that negotiating a uniform minimum price on carbon can have several desirable properties, including, especially, helping to internalize the global warming externality. To fully defend the relative "practicality" of what I am proposing would probably require a book, not an article. In any event, *this* article is not primarily about practical considerations of international negotiations. I leave that important task mostly to others.[9] However, I do want to mention just a few real-world considerations that have been left out of the model yet seem especially pertinent.

A key practical issue I am waving aside is just where in the production chain a carbon price should be collected. I think the presumption would be that the carbon price should be collected by the country in which the carbon dioxide is actually released into the atmosphere. One might try to argue that a carbon price should be collected downstream as close as possible to the point where the carbon is burned. But this would involve an impractically large number of collection points. It is much easier to collect the price upstream, at various chokepoints where the carbon is first introduced into the economy.[10]

Nothing in the model excludes side payments to help obtain an international agreement on harmonized national carbon prices. The transfer payments to lubricate compliance with harmonized national carbon prices could take the form of "contributions" from the developed countries that are earmarked for helping the developing countries finance low-carbon technologies.

A binding international agreement on a uniform carbon price presumably requires some serious compliance mechanism. To begin with, the carbon price must be observable. For enforcement, perhaps there is no practical alternative to using the international trading system for applying tariff-based penalties on imports from non-complying nations in the form of border-tax adjustments. Cooper (2010) has argued for an expansive interpretation whereby the internationally agreed charge on carbon emissions would be considered a cost of doing business, such that failure to pay the charge would be treated as a subsidy that is subject to countervailing duties under existing provisions of the World Trade Organization.[11]

I close by noting again that global warming is an extremely serious as-yet-unresolved externality problem. With the failure of a Kyoto-style quantity-based approach, the world

[9] See, e.g., Bodansky (2010) or Barrett (2005).

[10] This issues and its distributional consequences (including references to other literature) is discussed extensively in Asheim (2012).

[11] See also the discussion of the legality of such sanctions under WTO provisions in Metcalf and Weisbach (2009).

has seemingly given up on a comprehensive top-down global design, settling instead for sporadic national, sub-national, and regional measures. These partial measures seem far from constituting a socially efficient response to the global warming externality. Perhaps, as was previously suggested, the Kyoto-style quantity-based focus on negotiating emissions caps embodies a bad design flaw. The theoretical model here is indicating that negotiating a binding internationally-harmonized nationally-collected minimum price on carbon emissions might help to internalize the global warming externality. However, a more complete discussion of the policy implications of this theoretical finding is beyond the limited scope of this paper.

References

[1] Asheim, Geir B. (2012). "A Distributional Argument for Supply-Side Climate Policies." *Environ Resource Econ*, published online: 11 August 2012.

[2] Barrett, Scott (2005). *Environment and Statecraft: The Strategy of Environmental Treaty Making.* Oxford: Oxford University Press.

[3] Bodansky, Daniel (2010). *The Art and Craft of International Environmental Law.* Cambridge: Harvard University Press.

[4] Colman, A. M. (2006). "Thomas C. Shelling's psychological decision theory: Introduction to a special issue." *Journal of Economic Psychology*, 27: 603-608.

[5] Cooper, Richard N. (2010). "The Case for Charges on Greenhouse Gas Emissions." In Joe Aldy and Robert Stavins (eds), *Post-Kyoto International Climate Policy: Architectures for Agreement*, Cambridge University Press.

[6] Ellerman, A. Denny (2010). "The EU's Emissions Trading Scheme: A Prototype Global System?" In Joe Aldy and Robert Stavins (eds), *Post-Kyoto International Climate Policy: Architectures for Agreement*, Cambridge University Press.

[7] Goulder, Lawrence H., and Andrew R. Schein (2013). "Carbon Taxes vs. Cap and Trade: A Critical Review." Mimeo, August 2013.

[8] Hoel, Michael, and Larry Karp (2002). "Taxes vs. quotas for a stock pollutant." *Resource and Energy Economics* 24: 367-384.

[9] Libecap, Gary D. (2013). "Addressing Global Environmental Externalities: Transaction Costs Considered." NBER Working Paper 19501.

[10] Metcalf, Gilbert E., and David Weisbach (2009). "The Design of a Carbon Tax." *Harvard Environmental Law Review* 33.2: 499-556.

[11] Nordhaus, William D. (2007). "To tax or not to tax: Alternative approaches to slowing global warming." *Review of Environmental Economics and Policy* 1(1): 26-44.

[12] Nordhaus, William D. (2013). *The Climate Casino: Risk, Uncertainty, and Economics for a Warming World.* New Haven: Yale University Press.

[13] Pizer, William (1999). "Optimal Choice of Policy Instrument and Stringency under Uncertainty: the Case of Climate Change." *Resource and Energy Economics*, 12: 255-287.

[14] Schelling, Thomas C. (1960). *The Strategy of Conflict.* Harvard University Press.

[15] Weitzman, Martin L. (1974). "Prices vs. Quantities." *Review of Economic Studies* 41, 4: 477-491.