Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2010/13

An Experimental Contribution to the Theory of Customary (International) Law

Christoph Engel

MAX PLANCK SOCIETY

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2010/13

An Experimental Contribution to the Theory of Customary (International) Law

Christoph Engel

April 2010

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

An Experimental Contribution to the Theory of Customary (International) Law[1]

Christoph Engel

Abstract

In their majority, public international lawyers postulate that for a new rule of customary law to originate, two conditions must be fulfilled: there must be consistent practice, and it must be shown that this practice is motivated by the belief that such behaviour is required in law. Maurice Mendelson (Recueil des Cours 272 (1998) 155) has challenged this view. He believes that the majority view ignores the fundamentally incomplete nature of public international law. He claims that the new rule emerges because mere practice leads to convergent expectations. This paper uses data from student experiments with a linear public good to show that behaviour converges even absent verbal communication; that convergence is guided by mean contributions in the previous round, which serve as an implicit norm; that freeriding on this implicit norm is regarded as illegitimate; that cooperation can be stabilised at a high level if "reprisals" are permitted. Hence the mechanism of norm formation proposed by Maurice Mendelson is fully borne out by the experimental data.

JEL: C91, D03, D23, F53, H41, K33

[1] Helpful discussion with Konstantin Chatziathanasiou, Michael Kurschilgen, Alexander Morell and Niels Petersen and with the participants of the Ghent Seminar on Evolution of Law is gratefully acknowledged.

I. A Proper Concept of Customary (International) Law

Experimental public international law? Isn't it patent that such an endeavour is doomed to failure? Field experiments are ruled out in the first place; no state will agree to be treated at random, which would be necessary for identification. And isn't it obvious that a lab experiment with, say, student subjects is miles away from the problems states have to deal with? Yes and no. As with any experiment on law, external validity is an issue. In the discussion part of the paper I will address the challenge. Yet in the most fundamental respect, public international law is closer to the deliberately and radically context-free setting of the lab than any rule of domestic law. Public international law is more primitive, in the evolutionary sense. It lacks sovereignty. For sure, the addressees of public international law are sovereign states. Yet their sovereignty is confined to themselves. They have potentially unlimited rule making power internally, and in their dealings with each other they (grosso modo) respect each other's sovereign immunity. If they sign a treaty, this treaty can be said to rest in the combined sovereignties of the concluding parties. Yet above the level of states, there is no supreme authority that could ordain, let alone force, states to play by the rules, not even by those rules to which they have explicitly assented.

The absence of a supreme ruler explains why one source of law features prominently here that has almost died out in municipal law: customary law. In most textbooks of public international law, this source of law is treated much the same way as the sources of law in the introductory books on municipal law. In such texts, customary international law is tied back to a meta-rule that precisely defines the conditions under which a new rule of customary law comes into being. Textbooks typically list two conditions: sufficiently long and intense practice, and "opinio iuris". The latter requires that consistent practice results from the conviction of those contributing to it to be obliged in law (for a summary treatment see Treves 2009).

In his Hague lectures, Maurice Mendelson sharply criticises this approach (Mendelson 1998). For him, the textbook approach misses the categorical difference between municipal and public international law. The latter legal order is "semi-anarchic" (166), embryonic, and in a deep way incomplete. "Whilst modern domestic societies are characterised by highly centralised and compulsory systems of law-making and adjudication, not to mention enforcement, international society is not like that" (168). Therefore a "formalistic approach" (168) is misplaced. It is not possible to state in an abstract way the conditions that must be fulfilled for a new rule of customary international law to come into being (172). "The characteristic of this kind of law is that it is not just unwritten, it is informal" (172). "The customary process is in fact a continuous one, which does not stop when the rule has emerged […]. Even after the rule has 'emerged', every act of compliance will strengthen it, and every violation, if acquiesced in, will help to undermine it" (175). Customary international law rests on the conviction that "states should comply with the legitimate expectations of the international community", where the ambiguity of the term "expectation" is deliberate: "If, within a social group, people habitually behave in a certain way, then, particularly if others rely on the continuation of this conduct, the sentiment may develop within that society that one is obliged to continue so to act. In other words, a norm emerges from

what is normal […]. If the generality of states has regularly behaved in certain ways […], then a legitimate expectation arises that they will continue to do so" (185 f.).

These are testable propositions. Maurice Mendelson himself does not make a claim that is confined to international relations. He refers to a process "within a social group" (185). It therefore is meaningful to overcome the impossibility of testing states in the lab by studying the underlying social mechanism in an artificially created society with readily available subjects. To that end, this paper reanalyses a rich dataset composed of partly our own data, and partly data from structurally identical experiments run in labs all over the world.

The paper is in the spirit of those who have set out to analyse public international law in general (Keohane 2002; Van Aaken, Engel et al. 2008), and customary international law in particular, with the apparatus of rational choice theory. Such approaches assume that states are actors, that states (and not only individuals within states) have identifiable interests, and that states strive for realising their preferences, given the constraints resulting from international relations (Goldsmith and Posner 2005). Yet this paper does not explain the emergence of new rules of customary international law in game theoretic terms, as (Goldsmith and Posner 1999; Chinen 2001; Swaine 2002; Norman and Trachtman 2005; Norman and Trachtman 2008). While this explanation definitely has value, this paper stresses the implicit character of the norm generation process, and its evolutionary nature.

The paper is even closer to those who claim that state action is, at least partly, guided by the forces of "acculturation" (Goodman and Jinks 2004; Goodman and Jinks 2008). Yet these approaches explain norm emergence and norm compliance with a much richer set of contributing factors, both individualistic and social. While they distinguish acculturation and (more explicit, more intrusive) persuasion, they insist that acculturation is backed up by (social) sanctions, like shaming or public approval. By contrast, this paper treats sanctions as an additional explanatory variable, not as a necessary component of the underlying process. In a more sociological spirit, acculturation is understood as the process of becoming initiated to a group. By contrast, this paper stays as individualistic as possible. Since group composition is random, interaction is anonymous, and action is the only communication channel, one may wonder whether it makes sense to speak of culture in the first place. At any rate, the role of culture is a minimal as possible. This paper may thus be read as a radicalisation of the acculturation thesis: to the extent that one can show norms to even emerge in this radically decontextualised setting, they may *a fortiori* be expected to emerge in the considerably richer institutional setting of international relations.

II. Defining the Governance Problem

In legal textbooks, the sources of law are presented in a way that deliberately abstracts from their substance. When dealing with domestic law, this is a useful intellectual division of labour. The rules on rules deal with the formal conditions under which an act of legislature acquires validity.

Since states are sovereign, they are in principle free to give the new rule whatever contents the legislator deems fit. The constitution may prohibit certain rules, in particular through the protection of fundamental freedoms. Yet this possibility does not affect the abstract rules that define the sources of law.

If one follows Maurice Mendelson, this must be different with customary law. The emergence of a new rule rests on the conviction of those arguably participating in the informal process of creating it that a certain conduct is to be expected. What is to be expected, and therefore how a certain practice is to be interpreted, depends on the character of the problem the purported rule is meant to solve.

The quintessential social problem for the solution of which the emergence of norms is instrumental is a social dilemma. In the most general terms a dilemma is a situation in which individual and social rationality fall apart. What would be best for society is not in the best interest of the individual. In terms of game theory, this statement can be made precise and testable (for introductions see Baird, Gertner et al. 1994; Scharpf 1997), (for applications to international relations see Holzinger 2003; Sandler 2004). The experiments reported here all model the dilemma the same way. They implement a linear public good (for an overview of the experimental literature on these games, see Ledyard 1995; Zelmer 2003). This class of problems is frequent in international relations. Two classic examples are the maintenance of peace, and measures to combat climate change.

In all experiments, participants face the following payoff function:

$$\pi_i = e - c_i + \mu \sum_{j=1}^{n} c_j$$

Participant i has payoff π. In every round, she receives endowment e. She is free to keep it, or to invest all or part of her endowment in a public project. Investment has linear cost c_i and return $\mu < 1$. However, all contributions of all n members of the group are not only beneficial for herself, but also for all other group members, with $n\mu > 1$. Consequently, the group as a whole is best off if all participants invest their entire endowments. However, individually each group member makes the highest profit if only the remaining members contribute while she freerides. Therefore in the one-shot game, game theory predicts that all players contribute nothing. This prediction is not changed by the fact that the game has been repeated in all experiments. For the number of rounds was always announced in advance. In the last round, the prediction from the one-shot game obviously applies. Players holding standard preferences anticipate this and preempt being exploited in the last round by defecting themselves in the penultimate round. Through perfect anticipation, this step is repeated, so that in the model all defect right from the beginning (Selten 1978; Rosenthal 1981).

III. The Data

Table 1 summarises the data set. Four experiments exactly implement the game as presented in the previous paragraph, with parameters $e = 20, \mu = .4, n = 4$. In this literature, this design is called a voluntary contribution mechanism. The remaining five experiments allow for decentral punishment. What this means and which punishment technologies are implemented will be explained below. Three experiments are from our own lab (denoted MPI; for further detail see (Beckenkamp, Engel et al. 2009)). Four experiments have been run in London (denoted NIK, for further detail see (Nikiforakis 2008)). One experiment is from Rennes (denoted DEN, (Denant-Boèment, Masclet et al. 2007)). The final dataset consists of 16 identical experiments run in labs all over the world (denoted HER, (Herrmann, Thöni et al. 2008)). All but one experiment had 10 announced rounds; in the one exception, the game lasted 12 announced rounds. The dataset comprises a total of 14720 data points, collected from 1440 subjects interacting in 360 groups of four. In all but two experiments, the groups stayed together for the entire game (so-called partner design). In the remaining two experiments, groups were re-matched every round (so-called stranger design). All experiments were computerised, using the software zTree (Fischbacher 2007). In our own experiments, we invited subjects using the software ORSEE (Greiner 2004). In all experiments, participants interacted anonymously. Payoffs were paid out at the end of the game in real money.

game-type	matching	dataset	# obs.	T	P techn.
VCM	P	MPI	240	10	-
VCM	P	NIK	960	10	-
VCM	P	MPI	480	12	-
VCM	S	NIK	960	10	-
Pun	P	DEN	480	10	FG
Pun	P	MPI	240	10	FG
Pun	P	NIK	480	10	FG
Pun	P	HER	10400	10	1:3
Pun	S	NIK	480	10	FG

Table 1

Data Set

IV. What is the Problem and What is the Solution?

Figure 1 contrasts the aggregate experimental finding (the blue line) with the theoretical prediction (the green line). On average, in the beginning participants contribute much more than the theoretical expectation of zero. Contributions slowly but steadily decline. Yet as the error bars show, even at the end of the game, contributions are still significantly above the theoretical prediction.

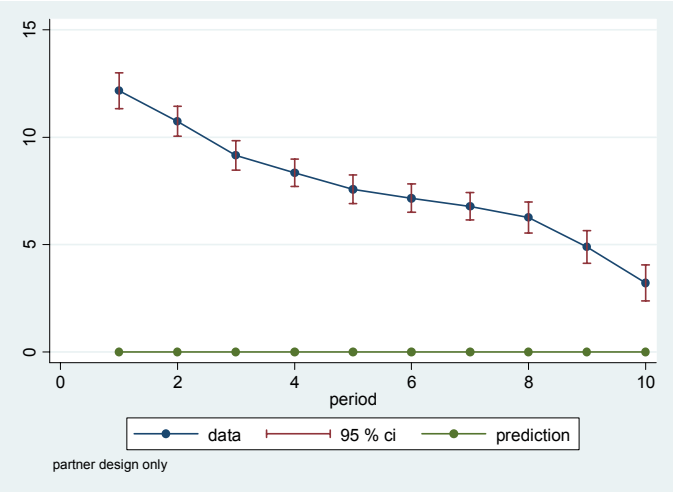

Figure 1

Development of Contributions in an Institution Free Setting

As always with experimental data, there is heterogeneity between groups. Yet as Figure 2 shows, the main finding is fairly robust. In no group, contributions are zero in the first period. Full contributions occur (groups 18, 19, 23, 412, 415), but eventually cooperation breaks down. Upward movements may occur for a limited number of periods, but a longer positive trend is very rare (see groups 416, 415, 418). The typical picture is substantial contributions in the beginning, and erosion over time.

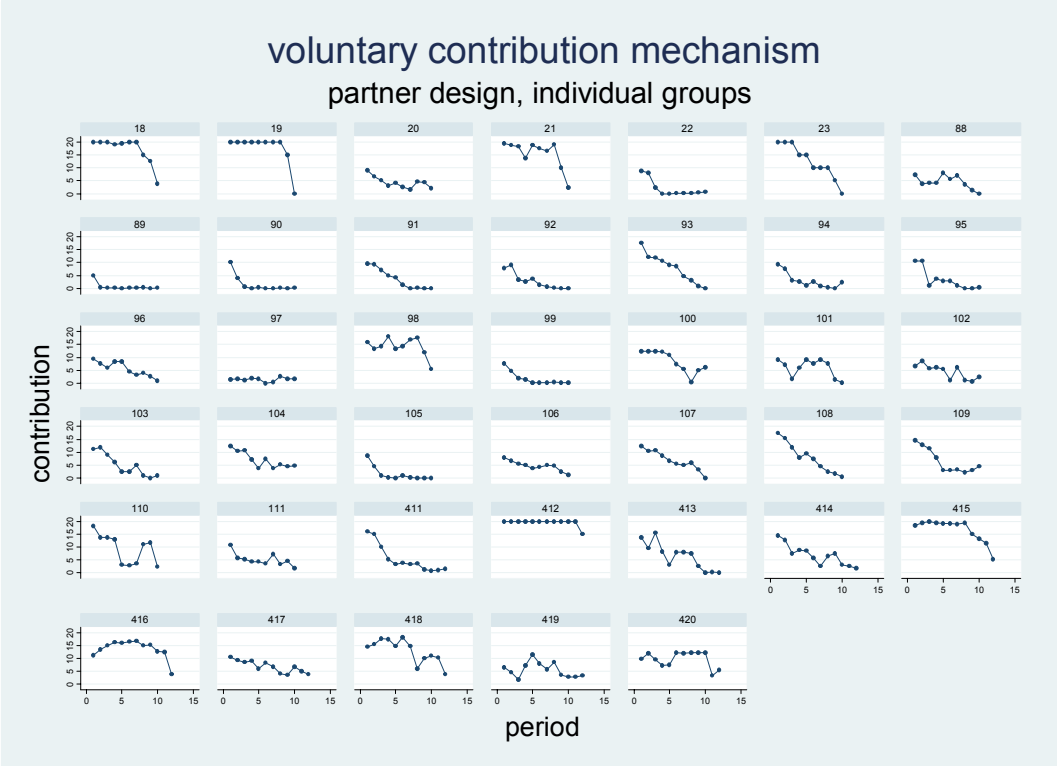

Figure 2

VCM: Individual Groups

Figure 3 looks at the initial period more closely. In this period, no more than 15.63 % of all 160 participants are in line with the theoretical prediction and contribute nothing to the joint project. By contrast, 32.5 % contribute their entire endowments. 70 % contribute at least half of the endowment.

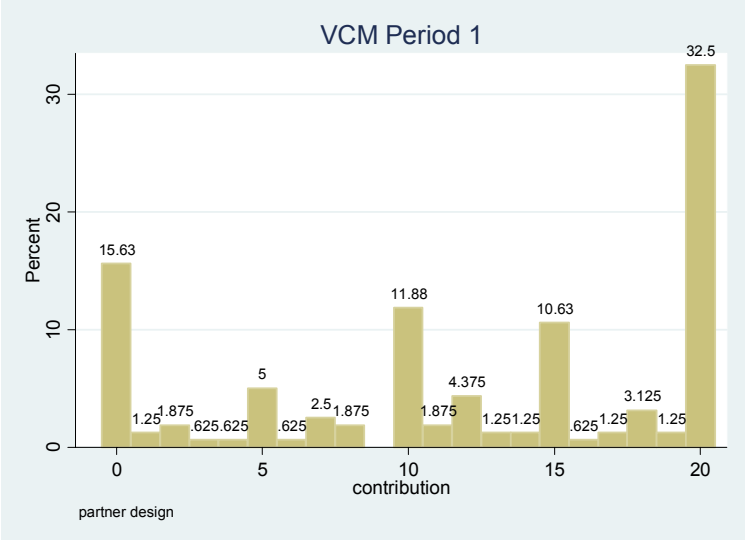

Figure 3

VCM: Period 1 Contributions

Given these findings, it is almost a philosophical question what is the problem and what is the solution. One may say that the problem is imperfect cooperation, but one may in as well say that imperfect defection is the solution.

Let us now explore the imperfection. As one directly sees in Figure 1, contributions decay over time. The negative time trend is statistically significant. In 39 of 40 independent groups, the mean change from one period to another is negative.[2] Table 2 provides a parametric estimate.[3]

[2] A one-sample signrank test of the nul hypothesis that this mean is zero rejects at $z = -5.498$, $p < .0001$.

[3] Parametric estimation of this data is demanding. Each participant decides repeatedly, which is why a fixed or random effects model is in order. Many participants contribute 20, quite a few contribute 0. Therefore the data is left and right censored, which calls for a Tobit model. Finally, since the groups of four stay together throughout the game, standard errors must be corrected for this relatedness. Unfortunately, there is no sandwich estimator for random effects Tobit models, which is why I bootstrap the models, with drawings at the level of groups. Finally, I perform the Hausman test on the mirror model that ignores clustering and censoring. In this model, the test turns out insignificant, which is why I am justified to use the more efficient random effects model.

contribution	
period	-1.571***
cons	14.459***
N	1600
left censored	551
right censored	266
p model	<.001

Table 2

Voluntary Contribution Mechanism: Time Trend

random effects Tobit, bootstrapped at the group level, 50 reps
*** p < .001, ** p < .01, * p < .05, + p < .1

As Figure 4 shows, the overall negative trend results from a negative balance. There are also upward movements, in particular in reaction to experiences from the first period. Yet downward movements are both more frequent and more pronounced. Over all periods, in all but one group there is a negative trend. From period 1 to 2, 8 from 40 groups move up, while 26 move down (and the remaining stay stable).

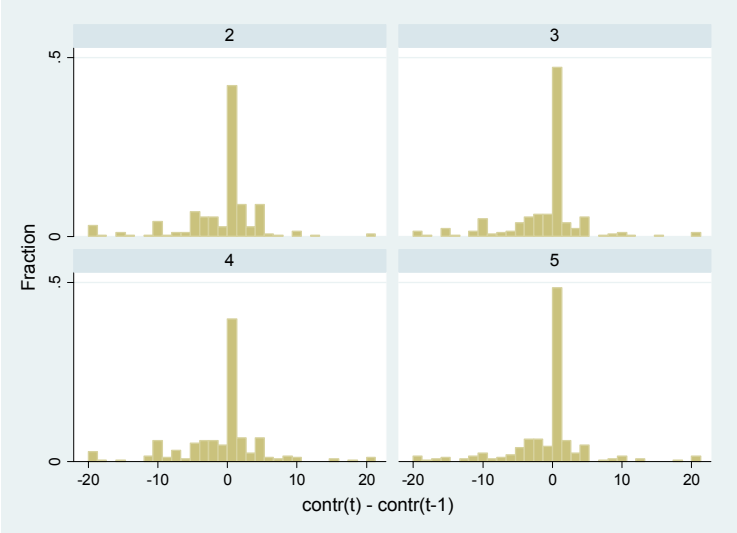

Figure 4

Time Trend in Initial Periods

V. Is There a Norm?

Thus far, we have only seen that there is a trend. There is a regularity. As Figure 1 shows, standard errors are relatively small, but they do not substantially reduce over time. In that sense, we cannot even claim to see convergence. Yet these are statements regarding the unconditional development. Maurice Mendelson has made a more ambitious claim: norms emerge by the very fact that mutual expectations match. In the experiments, verbal communication is excluded. The only way to get at expectations is studying behaviour. Specifically we can investigate how par-

ticipants react in the subsequent period to the experiences they have made in the previous period. If they had contributed less than the average and if they react by increasing their contributions, we can interpret this as an adjustment to the perceived expectation of others to make a higher contribution. Mendelson stresses that norms need not adjust upwards. If many do not play by what this player believes to be the rule, she may react by herself stopping to abide by it. In our setting, we do have a continuous variable, so that participants can also gradually reduce their own contributions in response to experienced frustration. As Figure 5 shows, both are indeed happening.

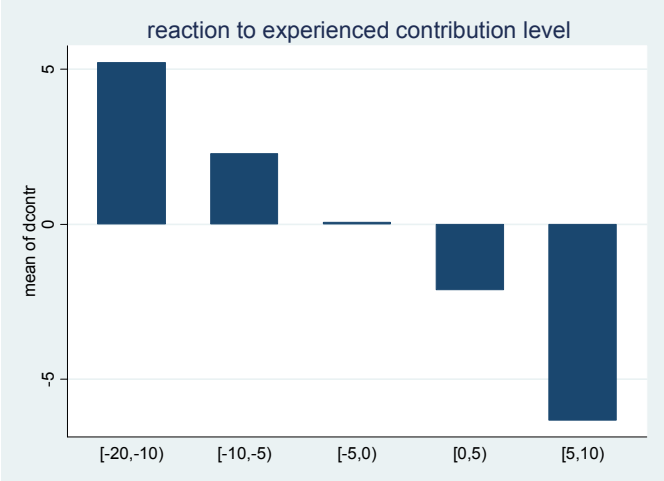

Figure 5

Contribution Changes, Conditional on Experienced Cooperation

In no group participants who had contributed in period t-1 *more* than the average of period t-2 on average increase their contributions in period t; in 38 of 40 groups they on average reduce their contributions.[4] By contrast, in 28 of 39 groups participants who had contributed in period t-1 *less* than the average of period t-2 on average increase their contributions in period t.[5] The result is confirmed by the parametric fixed effects models in Table 3. The negative constant reflects the fact that, overall, contributions decay over time. Note that, for those who contributed in t-1 less than the average in t-2, the independent variable is negative. Therefore the negative regressor implies that such participants increase their contributions, the more so the more they had been below the group average

[4] One sample signrank test, z = - 5.491, p < .0001.

[5] One sample signrank test of the nul that they do not change their contributions, z = 4.082, p < .0001.

.first differences of contributions	below average in previous period	above average in previous period
distance from average	-.508***	-.618***
cons	-1.162***	-1.530***
N	715	475
R^2 within	.1159	.0716
R^2 between	.0854	.3730
R^2 overall	.1185	.1429
p model	.0004	<.0001

Table 3

Reaction to Distance from Group Mean

fixed effects, clustered at group level
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

In opposition to the doctrinal mainstream, Maurice Mendelson urges public international lawyers to not require proof of *opinio iuris*. He believes the legitimacy of expectations results from mere practice. With our data, we can test whether mere practice suffices to generate both consistent behaviour and a normative expectation.

Opinio iuris critically presupposes discourse. Some states must claim that this is the law. Others must explicitly agree, or at least not object to the explicit claim. Our setting excludes explicit communication by design. If we nonetheless find action motivated by perceived legitimacy, we have shown Maurice Mendelson's claim to be true. We have an indirect measure for perceived legitimacy. Below we will be investigating how behaviour develops if participants are given a chance to express disapproval through costly punishment. We then will interpret decentral punishment as a (very embryonic) institutional intervention. At this point, we are only interested in the expressive function of punishment. In earlier periods, we cannot disentangle the "deontic" and the "consequentialist" functions of punishment. A player may punish because she hopes for a positive effect on contributions in subsequent periods, and exerts punishment effort as an investment into cooperativeness, in her group. Or she may punish because she is a retributionist and thinks a freerider deserves a sanction. Yet in the last period of the game, punishment can no longer be forward looking. If there is punishment, it must serve an expressive function. The punisher sends a signal of her discontent with the behaviour of another participant. If punishment in the last round is sensitive to the deviation from the contribution level in the penultimate round, punishment demonstrates that punishers regard negative deviations from the norm as unfair. They interpret the contribution level as a normative expectation.

First, Figure 6 demonstrates that in the last period punishment is still substantial. Actually, neither non-parametrically nor parametrically, there is a significant difference between the intensity of punishment in earlier and in the last period.[6] This shows that punishment cannot be exclusively instrumental.

[6] Mann Whitney, mean punishment in earlier periods vs. punishment in the final period, per group, N = 576, p = .1234; random effects Tobit, depvar: received punishment, indepvar: dummy = 1 if last period, coef. .283, p = .141, N = 12080, bootstrapped with drawings at the group level, 50 reps.

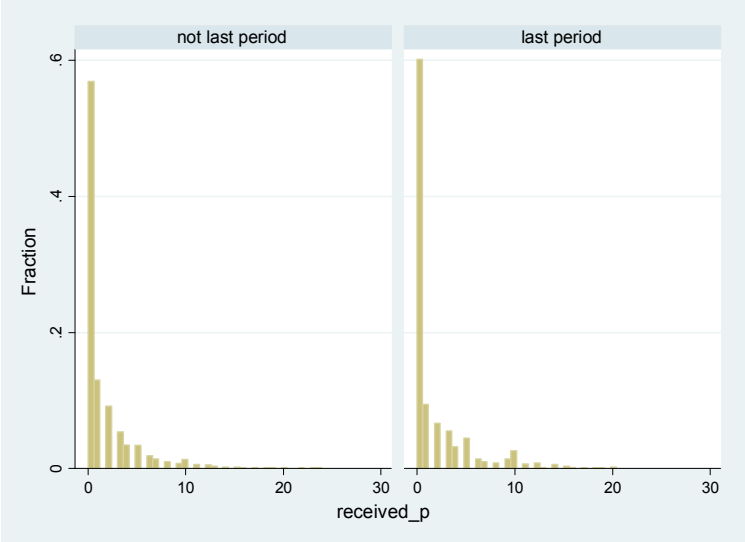

Figure 6

Intensity of Punishment Throughout and at the End of the Game

In the next step, Figure 7 shows that punishment in the final period is indeed sensitive to the degree by which the punishee deviates from the average contribution in this period, i.e. to the degree of freeriding. This finding too is statistically significant.[7] We do indeed establish the mechanism of norm evolution hypothesised by Maurice Mendelson.

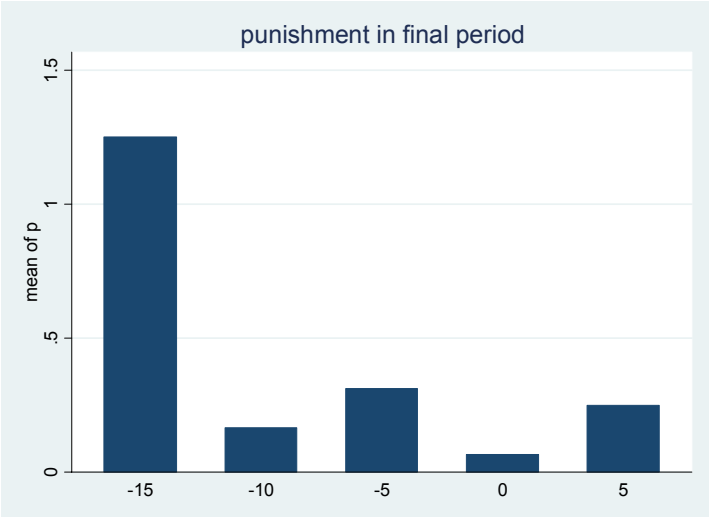

Figure 7

Sensitivity of Punishment in the Final Period to Degree of Freeriding

[7] Random effects Tobit, depvar: decision of participant a to punish participant b, in the last period of the game, N = 480, coef. (contr b – mean contr) -.271, p < .0001, cons -4.307, p <.0001, p model <.0001, Hausman test insignificant on mirror model that ignores censoring.

VI. Norms Are Context Contingent

From the concept of customary law suggested by Maurice Mendelson, it directly follows that the contents of customary rules is context contingent. If, at a certain point in time, say after the end of a large war, a vast majority of states believes time is ripe for a number of rules meant to make it more difficult to go to war, rules will emerge that would have been very unlikely to form in a different period of time. In principle, customary rules can also be contingent on geography. Admittedly, the typical rule of customary international law is universal. But it is undisputed that the geographic scope of customary international law may be more narrow. Then custom may emerge in some region, and not in others.[8] There are for instance a number of customary rules whose field of application is confined to Latin America, for instance with respect to asylum.

Context contingency can be tested in the lab, and using experimental data the concept of context contingency can be made more precise. Two related sources of contingency exist in the experimental data on voluntary contribution mechanisms. Not all experimental participants are equal. These idiosyncrasies interact to produce characteristics of each randomly composed experimental group.

Let us start with the second. In so doing, we operationalise context by past interaction patterns, within the group of which a participant happens to be a member. In the standard public goods experiment, those participants who are to form a group for the next 10 periods do not know each other's identity. Interaction is fully anonymous. The only channel of communication is action. Therefore the only possibility for forming a group is the contribution pattern. Actually according to the protocol used by the experiments reported in this paper, participants do not even learn the individual behaviour of other participants. All they see is their own payoff, from which they can deduct the average contribution in the respective period.[9]

Participants are very likely to use this information. Figure 8 demonstrates that one piece of information organises the data very well. The mean contribution in the first round is a very good predictor for contributions in later rounds. The degree of cooperativeness participants experience in the first round sets the stage. Participants gain a sense of what is feasible in this setting (more from Beckenkamp, Engel et al. 2009). Actually, in the graph most dots are below the line. This should not come as a surprise. Since Figure 1 shows that it is normal, in this game, for contributions to decay over time, the mean contribution in later rounds must typically be smaller than the mean contribution in the initial round. The interesting message is that most of these points are more or less in parallel to the line, as demonstrated by the regression line. Again, the graphical impression is corroborated by statistical analysis.[10] For a parametric test see Table 4.

---

8 ICJ Rep. 1950, 266, 277 – Asylum; Restatement (3rd) of the Foreign Relations Law of the U.S. § 102.

9 This too is different with punishment. Then each participant in each period sees how much each other group member has contributed. However group members are not identified across periods, and they of course remain anonymous.

10 Spearman's rho of mean contributions per group in periods 2-10 with average contributions, in this group, in period 1, .792, $p < .0001$.

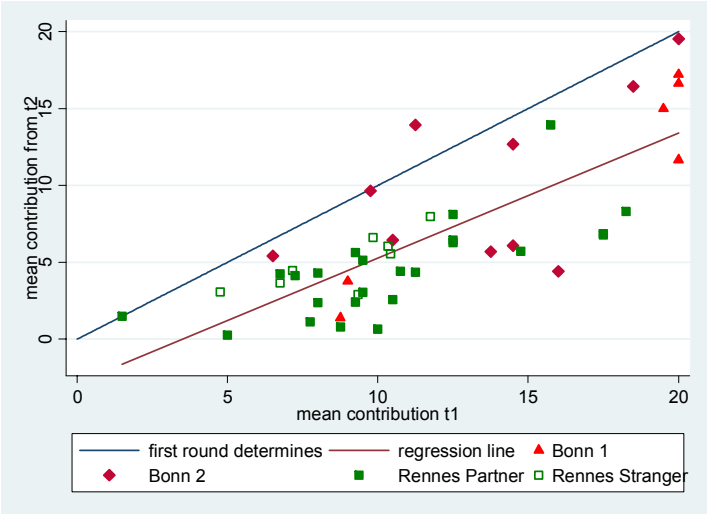

Figure 8

Effect of First Impressions

contribution in periods > 1	
mean contribution in period 1	1.388***
partner matching	-1.262
period	-1.372***
cons	-2.295^{+}
N	2384
p model	<.001

Table 4

Effect of First Impressions

random effects Tobit, bootstrapped at the group level, 50 reps
Hausman test on mirror model insignificant
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

Group heterogeneity does of course not fall from heaven. It is the product of the heterogeneity of group members. There are several possibilities for characterising group members. If one aims at explaining the group mean in the first period, unconditional first round contributions are the right measure. It however is even more revealing to consider conditional types. In principle, type should matter throughout the game. Hence if one were only interested in classification, one would want to use all data. Yet then type would be endogenous, and could no longer be used for explanation. I therefore only use the first three periods for classification. I form three groups: those who have in both periods been above the group mean of the respective previous period (whom I call "altruists"); those who have in both periods been below the group mean of the respective previous period (whom I call "freeriders"), and the intermediate group. I call them "conditional cooperators" since either they have just given the average from the previous period, or their contribution has oscillated around the group average of the previous period. Note that there are more direct measures of conditional cooperation (Fischbacher, Gächter et al. 2001; Fischbacher and Gächter 2010). Yet since these measures have not been employed in the experiments reported here, I must revert to the described proxy.

As Figure 9 highlights, players do indeed exhibit behavioural patterns throughout the game. Those whom I have classified as freeriders are more likely to give nothing in later rounds. They very rarely give more than 10 points. By contrast, those whom I have classified as altruists are least likely to give nothing, and they are fairly likely to contribute more than 10 points. In both respects, those whom I have classified as conditional cooperators are in the middle (but giving everything is also pronounced with them). Non-parametrically, the difference in means between the first and the two remaining groups is significant.[11]

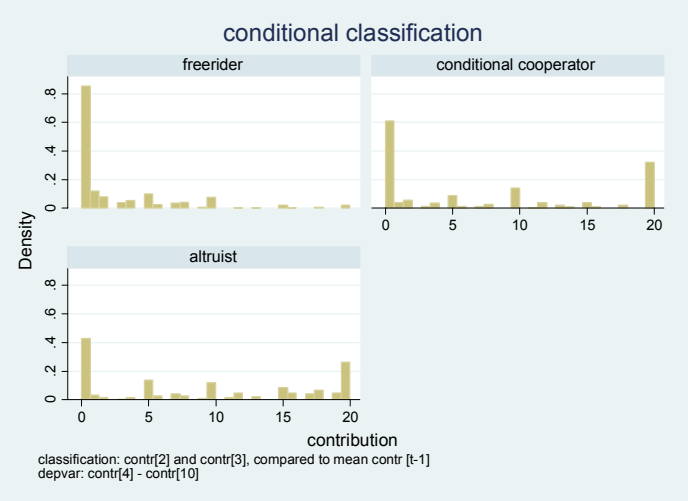

Figure 9

Effect of Player Type on Contributions in Later Rounds

Since any classification draws artificial lines, it is even safer to perform a classification free statistical test. The model of Table 5 directly works with the difference between this player's contributions in periods 2 and 3, compared to the average contribution, in her group, in the antecedent period. Independently, both the generosity of a subject in periods 2 and 3 significantly explain contributions in later rounds. Both coefficients are positive, indicating that cooperativeness early on is predictive of cooperativeness later in the game. Conversely, the model predicts that participants who were below the group average in early periods will also be below the average in later periods.

[11] Mann Whitney, freerider vs. conditional cooperator, N = 61, p = .0108; freerider vs. altruist, N = 55, p = .0114.

contribution in periods 4 – 10	
contr [2] – avcontr [1]	.414**
contr [3] – avcontr [2]	.619**
period	-1.735***
cons	16.843***
N	960
p model	<.0001

Table 5

Classification Free Test of Player Type Influence on Behaviour

random effects Tobit, bootstrapped at the group level, 50 reps
Hausman test on mirror model insignificant
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

VII. A More Demanding Norm

From the perspective of norm theory, we have all we need. There are behavioural regularities. These regularities guide behaviour. Both from above and from below, in an experiment participants adjust their contribution level to the mean contribution in their group, in the previous period. Participants do so since they interpret contributions in the previous period as an implicit norm. The contents of the norm is contingent on first impressions, which in turn result from the idiosyncratic type of the individuals that happen to form a group.

Yet from a policy perspective, the result is disappointing. On average, the longer their behaviour is guided by the implicit norm, the less groups are able to overcome their dilemma. Let us first show that a majority of participants feels this way. Figure 10 shows that all classes of participants (classified the same way as in Figure 9) sometimes contribute more than the average of the previous round. In conditional cooperators and altruists, such behaviour is even quite frequent.[12] This makes only sense if these participants hope that, at least in the long run, others will follow suit. Overcommitment can thus be interpreted as an investment into the establishment of a more demanding implicit norm. This implies that the implicit norm does not gradually decay, because nobody was willing to support a more ambitious norm. Quite a few participants are even happy to sacrifice some personal profit for the purpose. Yet in the totally institution free environment of the voluntary contribution mechanism, those in favour of a more stringent norm are not able to protect themselves against exploitation by freeriders.

[12] The difference between freeriders and conditional cooperators (Mann Whitney, N = 61, p = .0074) and between freeriders and altruists (Mann Whitney, N = 55, p = .0490) is statistically significant.

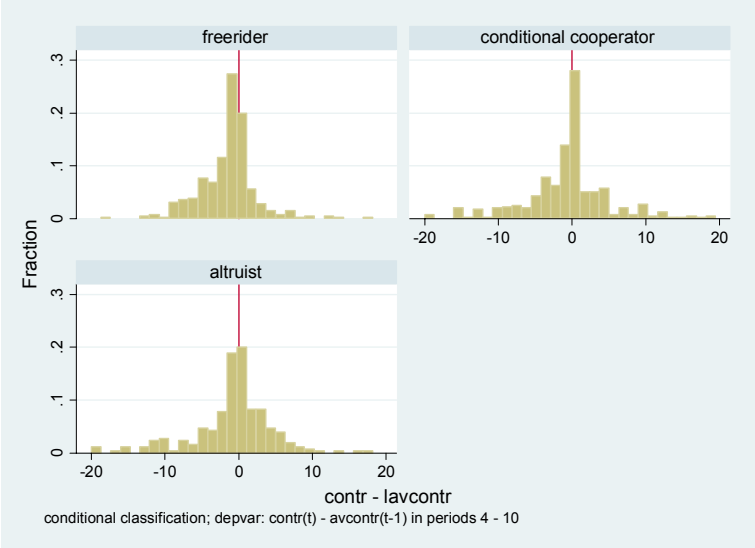

Figure 10

Overcommitment

Figure 11 demonstrates that this explanation indeed captures the essential driving force. While the red line repeats the earlier characteristic decay in the institution free environment, the blue line shows that cooperation stabilises at a fairly high level if participants are given the chance to punish each other after they have seen how much each of the other participants has contributed to the joint project, in the respective period.[13] Note that for a group of profit maximising participants, the punishment opportunity would be irrelevant. In the experiment, punishment is costly. Even if participants expect punishment (rightly) to induce a higher contribution level, each individual participant maximises her profit if she leaves it to the remaining group members to discipline freeriders, while she enjoys the higher period profit free of charge. The original dilemma thus repeats when it comes to disciplining the group (Heckathorn 1989). The difference in contributions is highly significant, even in a nonparametric test over means per group.[14] By a similar test, one establishes that, with punishment, the trend of contributions is significantly more positive.[15]

---

[13] Table 1 specifies the punishment technology. In four experiments, following (Fehr and Gächter 2000), the following cost function was used:

TABLE 2—PUNISHMENT LEVELS AND ASSOCIATED COSTS FOR THE PUNISHING SUBJECT

Punishment points p_i^j	0	1	2	3	4	5	6	7	8	9	10
Costs of punishment $c(p_i^j)$	0	1	2	4	6	9	12	16	20	25	30

By contrast (Herrmann, Thöni et al. 2008) use the linear technology originally introduced by (Fehr and Gächter 2002). According to this scheme, one punishment point destroys three points in the addressee.

[14] Mann Whitney, N = 336, p < .0001; the difference is also significant if we only compare results from the 22 groups of our own lab, p = .0063, i.e. if we test Figure 11.

[15] Mann Whitney, N = 22, p = .0008. The dependent variable is the coefficient of a random effects model that, separately for each group, regresses contributions on period, controlling for the endgame effect through an additional regressor for the final period. For comparability, this test too is confined to the partner design data from our own lab.

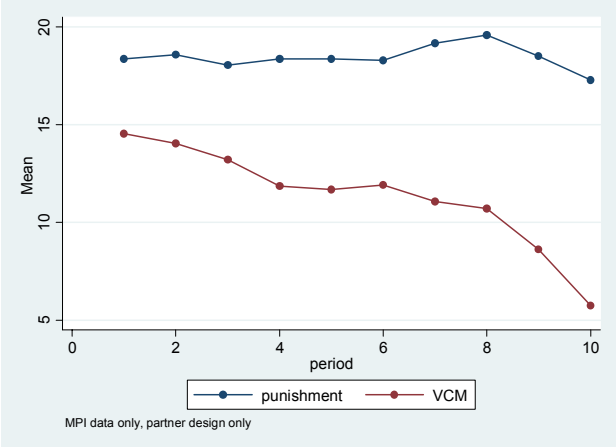

Figure 11

Contributions with and without Decentral Punishment

Note that the institution that proves so powerful in the lab is fairly close to the intervention states may face in international relations if they try to free ride on other states' efforts for the provision of a collective international good. As pointed out above, normally there is neither an authority for adjudication nor for enforcement to which a state could refer the dispute if it believes that another state violates an obligation under customary international law. All such states can do is themselves enforce the purported rule, through reprisals. Of course, the right to reprisals is sometimes abused. On the pretext that a disputed rule of customary law has been violated, a powerful state may itself violate an obligation under international law. Yet practically, reprisals are rare. This is understandable since an unwarranted reprisal is itself a violation of international law that may trigger countermeasures. Therefore what is a pecuniary cost in the lab chiefly is a risk in international relations.

VIII. Discussion

Experiments never map reality completely. Since the reality that rules of law are meant to govern is particularly rich, external validity tends to be a matter of concern in experiments on legal issues (also see Mendelson 1998:165-167). States are among the most aggregate corporate actors to be found on earth. By contrast, in the lab one studies the behaviour of isolated individuals. Public international law is an elegant tool for organising collectivities. One signature by the president binds 300 Mio Americans. Nonetheless many international conflicts engage a much larger number of actors than the four, or six, or eight members of an experimental group. While public international law is the offspring of centuries, in the experiments reported here participants interact for an hour or two. Consequently, historical conflict is absent by design. Injustices from the past do only matter if they have been inflicted a few minutes ago. In the experiment, the number of repetitions is announced, while in international relations states may not safely predict when a bilateral or multilateral relationship will terminate. Public international law deals with the essentials of this world, while our experimental subjects bargain over pennies. While in in-

ternational relations actors are almost always perfectly identified, the experiments provide perfect anonymity. This excludes reputation effects that are central in international relations. By the same token, communication is always an option between states, and one regularly seized. An important strand of international relations scholarship believes communication, or discourse as they tend to put it, is crucial for understanding the emergence of international norms (Keck and Sikkink 1998; Risse 1999; Risse 2000). In contrast, in the experiments verbal communication has been excluded. Actions have been the only communication channel. Even if explicit adjudication is the exception in international relations, it at least is a possibility, while it is excluded by design in the experiments. While history has given states very unequal opportunities, in the experiments opportunities are perfectly symmetric for all participants. Finally, in international relations ultimately no state can be prevented from going to war. Since the option of force goes unchecked, rights and obligations are never perfectly defined. By contrast, in the experiments each player's action space is precisely delineated.

All these differences certainly matter. Some could be relatively easily tested by new experiments. One could introduce uncertainty about the number of repetitions. Then theory would predict that cooperation is even easier to sustain (Aumann and Shapley 1994). One could lift anonymity and permit communication. One could make endowments or the action space asymmetric. One could introduce ambiguity to capture imperfectly defined property rights. One could raise stakes, for instance by playing the game in a developing country with a weak currency. Yet even if one were to do all that, important differences would remain. True corporate actors are next to impossible to implement in the lab (on experimental findings about the behaviour of corporate actors see Engel 2008). If one wants to maintain experimental control, true historical contingency is hard to implement as well. Adjudication would only be meaningful if the neat design of the opportunity structure were replaced by a sufficiently complex, and hence partly unpredictable, setting. Ultimately, one has to accept the trade-off. Experiments make it possible to solve the identification problem. If the experiment is properly designed, the arrow of causation is undisputed. Omitted variables can also largely be avoided. Yet these advantages have a price. Of necessity, the situation tested in the lab is much more naked than the situation in the field it is meant to explain. In the case of customary (international) law, this price is worth paying. For in the field, one may at best gain an intuition of what happens if a new rule of customary law emerges. By contrast, relying on the experiments reported in this paper one is able to precisely trace the evolutionary path. As Maurice Mendelson hypothesized, the essence of customary law is its evolutionary nature.

Literatur

AUMANN, ROBERT J. and LLOYD S. SHAPLEY (1994). Long Term Competition - A Game Theoretic Analysis. Collected Papers I. Robert J. Aumann. Cambridge, MIT Press**:** 395-409.

BAIRD, DOUGLAS G., ROBERT H. GERTNER and RANDAL C. PICKER (1994). Game Theory and the Law. Cambridge, Mass., Harvard University Press.

BECKENKAMP, MARTIN, CHRISTOPH ENGEL, ANDREAS GLÖCKNER, BERND IRLENBUSCH, HEIKE HENNIG-SCHMIDT, SEBASTIAN KUBE, MICHAEL KURSCHILGEN, ALEXANDER MORELL, ANDREAS NICKLISCH, HANS-THEO NORMANN and EMANUEL TOWFIGH (2009). Beware of Broken Windows! First Impressions in Public-good Experiments http://ssrn.com/abstract=1432393.

CHINEN, MARK A. (2001). "Game Theory and Customary International Law. A Response to Professors Goldsmith and Posner." Michigan Law Review **23**: 143-189.

DENANT-BOÈMENT, LAURENT, DAVID MASCLET and CHARLES NOUSSAIR (2007). "Punishment, Counter-Punishment and Sanction Enforcement in a Social Dilemma Experiment." Economic Theory **33**: 145-167.

ENGEL, CHRISTOPH (2008). The Behaviour of Corporate Actors. A Survey of the Empirical Literature

FEHR, ERNST and SIMON GÄCHTER (2000). "Cooperation and Punishment in Public Goods Experiments." American Economic Review **90**: 980-994.

FEHR, ERNST and SIMON GÄCHTER (2002). "Altruistic Punishment in Humans." Nature **415**: 137-140.

FISCHBACHER, URS (2007). "z-Tree. Zurich Toolbox for Ready-made Economic Experiments." Experimental Economics **10**: 171-178.

FISCHBACHER, URS and SIMON GÄCHTER (2010). "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments." American Economic Review **100**: 541-556.

FISCHBACHER, URS, SIMON GÄCHTER and ERNST FEHR (2001). "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." Economics Letters **71**: 397-404.

GOLDSMITH, JACK L. and ERIC A. POSNER (2005). The Limits of International Law. Oxford ; New York, Oxford University Press.

GOLDSMITH, JACK and ERIC A. POSNER (1999). "A Theory of Customary International Law." University of Chicago Law Review **66**: 1113-1177.

GOODMAN, RYAN and DEREK JINKS (2004). "How to Influence States. Socialisation and International Human Rights Law." Duke Law Journal **54**: 621-704.

GOODMAN, RYAN and DEREK JINKS (2008). "Incomplete Internalization and Compliance with Human Rights Law." European Journal of International Law **19**: 725-748.

GREINER, BEN (2004). An Online Recruiting System for Economic Experiments. Forschung und wissenschaftliches Rechnen 2003. Kurt Kremer und Volker Macho. Göttingen**:** 79-93.

HECKATHORN, DOUGLAS D. (1989). "Collective Action and the Second-Order Free-Rider Problem." Rationality and Society **1**: 78-100.

HERRMANN, BENEDIKT, CHRISTIAN THÖNI and SIMON GÄCHTER (2008). "Antisocial Punishment Across Societies." Science **319**: 1362-1367.

HOLZINGER, KATHARINA (2003). Transnational Common Goods. Strategic Constellations, Collective Action Problems, and Multi-Level Provision.

KECK, MARGARET E. and KATHRYN SIKKINK (1998). Activists Beyond Borders. Advocacy Networks in International Politics. Ithaca, N.Y., Cornell University Press.

KEOHANE, ROBERT O. (2002). "Rational Choice Theory and International Law. Insights and Limitations." Journal of Legal Studies **31**: 307-319.

LEDYARD, JOHN O. (1995). Public Goods. A Survey of Experimental Research. The Handbook of Experimental Economics. J.H. Kagel und A.E. Roth. Princeton, NJ, Princeton University Press**:** 111-194.

MENDELSON, MAURICE H. (1998). "The Formation of Customary International Law." Recueil des Cours **272**: 155-410.

NIKIFORAKIS, NIKOS S. (2008). "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" Journal of Public Economics **92**: 91-112.

NORMAN, GEORGE and JOEL P. TRACHTMAN (2005). "The Customary International Law Game." American Journal of International Law **99**: 541-580.

NORMAN, GEORGE and JOEL P. TRACHTMAN (2008). "Measuring the Shadow of the Future. An Introduction to the Game Theory of Customary International Law." University of Illinois Law Review: 127-154.

RISSE, THOMAS (1999). "International Norms and Domestic Change. Arguing and Communicative Behaviour in the Human Rights Area." Politics & Society **27**: 529-559.

RISSE, THOMAS (2000). ""Let's Argue!". Communicative Action in World Politics." International Organization **54**: 1-39.

ROSENTHAL, ROBERT W. (1981). "Games of Perfect Information, Predatory Pricing and the Chain Store Paradox." Journal of Economic Theory **25**: 92-100.

SANDLER, TODD (2004). Global Collective Action. Cambridge, England ; New York, Cambridge University Press.

SCHARPF, FRITZ WILHELM (1997). Games Real Actors Play. Actor-Centered Institutionalism in Policy Research. Boulder, Colo., Westview Press.

SELTEN, REINHARD (1978). "The Chain Store Paradox." Theory and Decision **9**: 127-159.

SWAINE, EDWARD T. (2002). "Rational Custom." Duke Law Journal **52**: 559-627.

TREVES, TULLIO (2009). "Customary International Law." Max Planck Encyclopedia of Public International Law: 1-20.

VAN AAKEN, ANNE, CHRISTOPH ENGEL and TOM GINSBURG (2008). "Public International Law and Economics." University of Illinois Law Review: 1-436.

ZELMER, JENNIFER (2003). "Linear Public Goods. A Meta-Analysis." Experimental Economics **6**: 299-310.