European Sociological Review, 2016, Vol. 32, No. 3, 439–451
doi: 10.1093/esr/jcw014
Advance Access Publication Date: 13 April 2016
Original Article

OXFORD

The Sanctioning Dilemma: A Quasi-Experiment on Social Norm Enforcement in the Train

Wojtek Przepiorka[1,2,*] and Joël Berger[3,4]

[1]Department of Sociology, Utrecht University, Utrecht 3584 CH, The Netherlands, [2]Nuffield College, New Road, Oxford OX1 1NF, UK, [3]Department of Sociology, University of Groningen, 9712 TG Groningen, The Netherlands and [4]ETH Zurich, CH-8092 Zurich, Switzerland

*Corresponding author. Email: w.przepiorka@uu.nl

Submitted July 2015; revised March 2016; accepted March 2016

Abstract

Numerous laboratory experiments have established peer-sanctioning as an important driver of norm compliance and cooperation in human groups. However, systematic evidence of peer-sanctioning occurring in the field is still rare. Here we present results from a quasi-experimental field study investigating the enforcement of the silence norm in the train. We let a confederate play music on his/her mobile phone in an open-plan train car and measure the time until a negative sanction occurs (if any). The silence norm is enforced in 45 of 90 cases, enforcement rates do not differ across silent- and non-silent-area cars, and the more passengers are in a car, the more likely is the silence norm enforced. Passengers' propensities to enforce the silence norm are in line with predictions derived from the asymmetric volunteer's dilemma (VOD). The higher a passenger's net benefit from enforcing the silence norm is, the more likely is the passenger to negatively sanction the norm breaker. Our findings extend the validity of results from laboratory experiments which conceive the second-order free-rider problem as a VOD.

Introduction

Social norms and their enforcement through positive and negative sanctions are indispensable for social cohesion and the functioning of societies. While many social norms are formalized in terms of paragraphs in legal codes, far from all social norms can be formally defined and even fewer enforced at all times by a legitimate authority. Although not unheard of, most people are reluctant to instantly call the police if someone is smoking in a non-smoking area, listening to loud music in public transport, jumping the queue in the supermarket, dressing inappropriately at a wedding, or free-riding on a group project at work. Such norm violations are, if at all, negatively sanctioned by the norm breakers' peers.[1] Peer-sanctioning comes in many forms; everyday norm violations can be sanctioned by disapproving words, looks, or gestures, by negative gossip and ostracism, or acts of aggression (Brauer and Chekroun, 2005; Guala, 2012; Feinberg, Willer and Schultz, 2014). But why do people engage in the informal sanctioning of their deviant peers—even at the risk of receiving an aggressive response?

In most instances, peer-sanctioning can be explained by the fact that the norm breaker and the sanctioner are going to meet again in the future (e.g. next-door neighbours, colleagues at work, individuals and organizations with a reputation to lose); that is, when the benefits of establishing a cooperative environment outweigh the costs of upholding the sanctioning threat in the long run (Gächter, Renner and Sefton, 2008; Horne, 2009).

However, in many instances of everyday norm violations, it is *a priori* unlikely that the norm breaker and the affected parties will encounter each other again. Yet, many of those who are affected by the norm violation object and sanction the norm breaker. Reciprocity and other-regarding preferences have been suggested as a proximate explanation for the sanctioning of norm breakers in one-time-only encounters. Theoretical considerations (Gintis, 2000) as well as empirical findings from laboratory experiments (Ostrom, Walker and Gardner, 1992; Fehr and Gächter, 2002) corroborate that many humans are indeed inclined to 'sacrifice resources for rewarding fair and sanctioning unfair behavior *even if this is costly and provides neither present nor future material rewards*' (Fehr, Fischbacher and Gächter, 2002: p. 3, emphasis in original).

These findings have been contested in their empirical and methodical validity (Levitt and List, 2007; Dreber et al. 2008; Herrmann, Thöni and Gächter, 2008; Nikiforakis, 2008; West, El Mouden and Gardner, 2011; Guala, 2012; Krasnow et al., 2012). However, only relatively recently have experimental social psychologists (e.g. Diekmann et al., 1996; Chekroun and Brauer, 2002; Brauer and Chekroun, 2005) and experimental economists started addressing issues of external validity by investigating peer-sanctioning in the field (Balafoutas and Nikiforakis, 2012; Balafoutas, Nikiforakis and Rockenbach, 2014). For example, Balafoutas and Nikiforakis (2012) conduct a field experiment to study the sanctioning of norm breakers in a large subway station in Athens. A confederate violated the 'do not litter in public places' norm and the 'stand right walk left on the escalator' norm 150 times each. Their results show that the no-litter norm was enforced in 4 per cent of the cases and the escalator norm was enforced in 19.3 per cent of the cases (for replications regarding the no-litter norm and the escalator norm in other cities, see Berger and Hevenstone, 2016, and Wolbring, Bozoyan and Langner, 2013, respectively). Based on potential sanctioners' statements in a follow-up questionnaire, Balafoutas and Nikiforakis (2012) conjecture that the low enforcement rate of the no-litter norm is due to potential sanctioners being afraid of retaliation. That is, since the no-litter norm is commonly better known than the escalator norm, the violation of the no-litter norm is a credible sign of the norm breaker's intentionality, anti-social preferences, and thus, his or her higher propensity to retaliate sanctioning. Ultimately, the relatively low sanctioning rates reported in these field studies cast further doubt on the external validity of results regarding peer-sanctioning obtained in laboratory experiments. We provide a comprehensive review of (quasi) experimental field studies on peer-sanctioning in the Supplementary Material.

We contribute to this literature by pointing out an understudied aspect of peer-sanctioning. Most studies investigating peer-sanctioning in the laboratory or in the field have neglected the strategic nature of many sanctioning situations. Like many scholars theorizing about social norms did before us (Ullmann-Margalit, 1977; Axelrod, 1986; Coleman, 1990; Ostrom, 2000; Horne, 2001), we start from the assertion that the violation of a social norm creates negative externalities for a group of people, and thus, the enforcement of the norm creates a (second-order) public good (Yamagishi, 1986; Heckathorn, 1989). Moreover, in many instances of everyday norm violations, a collective demand for negative sanctions is created that can be satisfied by one actor alone. And if the benefits outweigh the costs of producing the second-order public good, it can be in the actor's self-interest to produce it for the entire group (Olson, 1971 [1965]; Roberts, 2013).

We investigate the enforcement of the silence norm among passengers who happen to sit in the same open-plan train car. In 90 instances, we let confederates play annoying music on their mobile phone and measure the time until a sanction occurs (if any) as well as relevant contextual variables and passengers' characteristics. In a first study, comprising 31 interventions with a male confederate, we merely investigate the prevalence of norm enforcement. In a second study, comprising 59 interventions with female confederates, we also systematically vary whether the intervention takes place in a silent-area car or in a non-silent-area car.

Theory and Hypotheses

The violation of the silence norm in an open-plan train car constitutes the first-order free-rider problem. The quiet environment can be conceived of as a special case of a common pool resource (CPR; e.g. Ostrom et al., 1992). A person behaving loudly over-extracts the CPR at a cost for those who do not. The negative externality produced by the norm breaker manifests itself in other passengers' increased costs of following activities that require a certain degree of silence. Thus, sanctioning the norm breaker re-establishes a quiet environment for the remaining duration of the train ride and, as such, produces a public good. However, the sanctioning of the norm breaker is costly and therefore subject to a (second-order) free-rider problem (Yamagishi, 1986; Heckathorn, 1989). In what follows, we argue that this second-order free-rider problem can be modelled with

the volunteer's dilemma (VOD; Diekmann, 1985; Raihani and Bshary, 2011).

Diffusion of Responsibility in the Symmetric VOD

The VOD is a step-level public good game where only one actor's contribution is necessary and sufficient to produce the public good for the entire group (Palfrey and Rosenthal, 1984; Diekmann, 1985). In our case, sanctioning the norm breaker to re-establish silence for all passengers in the train car constitutes the (second-order) public good. More formally, a public good of value $\sum U_i$ for a group of size $n \geq 2$ is produced by a single actor i choosing to sanction the norm breaker at a cost K_i, where $U_i > K_i > 0 \ \forall \ i$. The public good is not provided if all actors choose not to sanction the norm breaker, and there is a welfare loss if more than one actor sanctions the norm breaker. The VOD thus has n welfare maximizing, pure strategy Nash equilibria, in which one passenger sanctions the norm breaker while all other $n-1$ passengers do not. Table 1 presents the pay-off structure of the VOD from passenger i's perspective.

The social dilemma comprised in the VOD arises from the fact that, without communication, it is difficult for a group of passengers to tacitly agree on which one of them should sanction the norm breaker. Although the benefits outweigh the costs of sanctioning (i.e. $U_i > K_i > 0 \ \forall \ i$), free-riding on another passenger's sanction is even more beneficial. As a consequence, the entire group may end up suffering from the negative externalities produced by the norm violation, while waiting for someone else to sanction the norm breaker. Assuming the *symmetric* VOD, where all passengers have the same benefits from and costs of sanctioning the norm breaker (i.e. $U_i = U_j$ and $K_i = K_j \ \forall \ i \neq j$), this diffusion of responsibility effect (Darley and Latané, 1968; Diekmann, 1985) can be derived from the mixed strategy equilibrium (MSE).

$$p_i^* = 1 - \sqrt[n-1]{K_i/U_i} \tag{1}$$

In the MSE, each passenger i sanctions the norm breaker with a certain probability p_i^*. Furthermore, with $q_i^* = 1 - p_i^*$, we can calculate the probability p^* that at least one passenger will sanction the norm breaker and the second-order public good will be produced (see Diekmann, 1985, for the derivation of equations 1 and 2).

$$p^* = 1 - \prod_{i=1}^{n} q_i^* \tag{2}$$

Consistent with the diffusion of responsibility effect, both p_i^* and p^* are decreasing in n, the size of the group. Based on our theoretical argument thus far, we can derive our first hypothesis:

H1: The larger a group of passengers sitting in the same train car is, the less likely will the silence norm be enforced.

Tacit Coordination in the Asymmetric VOD

Recall, however, that H1 derives from the MSE of the symmetric VOD, which assumes the same benefits and costs for all passengers. In the sanctioning situation under scrutiny, this assumption is likely to be violated. First, some passengers try to make the best of the time they spend in a train by, for instance, working or reading, whereas other passengers simply enjoy looking out of the window, dozing, consuming media (usually using their earphones), or chatting with others (possibly on the phone). Thus, the sanctioning of the norm breaker will produce a greater benefit for those passengers who require a quiet environment to follow their activities. Secondly, the costs arise from the actual act of sanctioning the norm breaker. That is, getting up, approaching the norm breaker, making an assertive statement that the music should be turned down, and possibly facing an aggressive response, all sum up to an individual's total cost of sanctioning. While male passengers derive on average the same benefits from enforcing the silence norm as female passengers, they are likely to have lower costs. Men are on average taller and more aggressive than women, and there are strong stereotypes describing man as more dominant and women as submissive (Lueptow, Garovich-Szabo and Lueptow, 2001). In the presence of hegemonic gender beliefs, male passengers will be believed to have lower sanctioning costs, because

Table 1. The volunteer's dilemma

	Number of other passengers who sanction the norm breaker			
Passenger *i*'s choice	0	1	…	*n* - 1
sanction norm breaker	$U_i - K_i$	$U_i - K_i$	$U_i - K_i$	$U_i - K_i$
don't sanction norm breaker	0	U_i	U_i	U_i

of their deterring effect on the norm breaker, and thus will more likely sanction the norm breaker (Ridgeway and Correll, 2004). Thirdly, irrespective of their activity and gender, passengers sitting closer to the source of noise will be more disturbed and, therefore, the sanctioning of the norm breaker will generate a larger benefit for them. Moreover, sitting closer to the norm breaker also reduces the costs of sanctioning as getting up to approach the norm breaker might not be necessary. Consequently, sitting closer to the source of noise, a passenger will more likely believe that passengers sitting further away will be less likely to sanction the norm breaker and thus he or she will more likely sanction the norm breaker himself or herself. Passengers sitting further away will have the corresponding beliefs and act accordingly.

The set of observable passenger attributes, i.e. passengers' distance to the source of noise and their activity and gender, create heterogeneity in passengers' net benefits from sanctioning the norm breaker. This heterogeneity can be accounted for with the asymmetric VOD (Diekmann, 1993), where $U_i \neq U_j$ and/or $K_i \neq K_j \; \exists \; i \neq j$. Based on the *asymmetric* VOD, it has been shown theoretically (Diekmann, 1993; He, Wang and Li, 2014) and empirically (Przepiorka and Diekmann, 2013; Diekmann and Przepiorka, 2015) that the person with the largest net benefit from norm enforcement will be the most likely to sanction the norm breaker (see also Brauer and Chekroun, 2005). Correspondingly, our next hypotheses are as follows:

H2-1: The closer a passenger sits to the source of noise, the more likely will this passenger enforce the silence norm.

H2-2: The more silence a passenger's activity requires, the more likely will this passenger enforce the silence norm.

H2-3: A male passenger will be more likely to enforce the silence norm than a female passenger.

Under asymmetric conditions, the diffusion of responsibility effect hypothesized under H1 will be less likely to occur. Passengers' perceivable differences in net benefits from norm enforcement facilitate the group's tacit coordination on the passenger with the highest net benefit to sanction the norm breaker (Przepiorka and Diekmann, 2013). Thus, in asymmetric situations, group size will have less bearing on the probability that a sanction occurs. H1 and H2 can thus be seen as putting two alternative models—the symmetric and the asymmetric VOD, respectively—to an empirical test.

Decreasing Benefits in the Volunteer's Timing Dilemma

Both the symmetric and asymmetric VOD are simultaneous move games, where all actors make their decisions at the same time, without knowing what other group members decide. However, the sanctioning situation in the train is dynamic in that all passengers can observe all other passengers' moves. That is, over time, passengers can update their beliefs about other passengers' propensity to sanction the norm breaker and refrain from sanctioning the norm breaker after someone else did. A dynamic version of the VOD was first described by Bliss and Nalebuff (1984). Weesie (1993) suggested a similar dynamic conceptualization of the VOD and called it the volunteer's timing dilemma (VTD). In both strands of literature, it is assumed that the benefits of establishing the (second-order) public good decrease over time. In the sanctioning situation in the train, the benefits of enforcing the silence norm decrease with the train approaching its destination. At the same time, the costs of enforcing the silence norm stay constant. We can therefore state our next hypothesis.

H3: The closer the train is to its destination, the less likely will the silence norm be enforced.

Moreover, H1 and H2 can also be derived from the symmetric and asymmetric VTD, respectively, but now with the timing of norm enforcement as the dependent concept (Weesie, 1993). For instance, H1 and H2-1 can be rephrased as follows: The larger a group of passengers sitting in the same train car is, the *later* will the silence norm be enforced; the closer a passenger sits to the source of noise, the *earlier* will this passenger enforce the silence norm. Given that each intervention lasts less than 5 minutes, and assuming that passengers form accurate beliefs about other passengers' sanctioning propensity *mainly* based on these other passengers' observable characteristics, we treat the two types of hypotheses H1 and H2 as equivalent. In the Results section, we estimate both models with the probability and timing of norm enforcement as dependent variables.

Norm Salience

Most people, at least in Switzerland, where the two studies were conducted, would expect most others to agree that overly loud behaviour in an open-plan train car is inappropriate and should be negatively sanctioned. However, passengers' beliefs might be different in trains in which there is an explicit distinction between silent-area cars and non-silent-area cars. In these trains, passengers in a silent-area car should feel more entitled

to reprimand someone breaking the silence norm than passengers in a non-silent-area car.

H4: The silence norm will be more likely enforced in a silent-area car than in a non-silent-area car.

However, based on what Balafoutas and Nikiforakis (2012) conjecture based on their findings (see above), we might also expect the opposite. Since the breaking of the silence norm in the silent-area car, where this norm is made explicit by visible signs in the car, would be indicative of the norm breaker's retaliation potential, passengers in a silent-area car should feel more reluctant to enforce the silence norm than passengers in a non-silent-area car.

H4a: The silence norm will be less likely enforced in a silent-area car than in a non-silent-area car.

Materials and Methods

Trains provide an ideal setting for conducting quasi-experiments (Shadish, Cook and Campbell, 2002), as they restrict passengers in their action space for some time and, at the same time, allow for careful measurement and controlled intervention (Levitt and List, 2007). In this section, we detail the design and procedures of the two studies we conducted to test our hypotheses. Since the two studies differ in a few but important respects, we start with an in-depth description of study 1 and then only point out the differences between the first and second study. In the last part of this section, we give a brief description of the data that we collected.

Procedures and Design

The first study was conducted in the intercity (IC) train between Zurich and Bern. This train comprises non-silent-area cars only, takes 58 minutes between destinations, and has no stops in between. In total, 14 train rides were taken, seven from Zurich to Bern and seven in the opposite direction. Between one and four interventions per train ride and 35 interventions in total were recorded. The first two train rides were used to test the intervention. We discarded four out of six interventions recorded on these two rides because they differ from subsequent interventions in the volume at which the music was played. In total, 31 valid interventions were conducted, and data on 204 passengers were collected. The number of interventions that could be recorded on one ride depended on how crowded the train was and therefore, on whether the experimenters could find an opportunity to carry out the intervention (see below). As mentioned earlier, the benefits from sanctioning a norm violation decrease with the train approaching its destination, whereas the costs stay constant. To avoid the costs of enforcing the silence norm from exceeding the benefits and thus allowing the sanctioning situation to fail to comprise a VOD, no interventions were started in the last 15 minutes of the ride. The interventions were conducted by two experimenters: one norm breaker, who violated the silence norm by playing a song on his mobile phone at an annoying volume, and one observer, who recorded the contextual variables, passenger characteristics and passengers' reactions to the norm violation. It was always the same person who was the norm breaker or the observer.

The following procedures were followed (also see Figure 1): The norm breaker (X) enters the train car and takes a seat in an empty compartment. Then, the observer (O) enters the car and takes a seat sufficiently distant from the norm breaker in order not to interfere with the intervention. The observer takes about 5 min to record the observable characteristics of the other passengers present in the car (e.g. passengers P1 through P6 in Figure 1) on a prepared form (one version of the form is presented in Supplementary Figure SA1). The other passengers' gender, estimated age, activity, and their

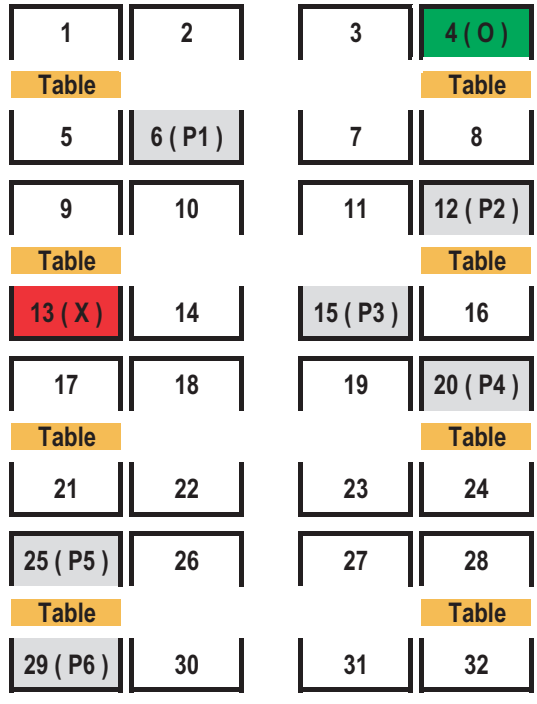

Figure 1. Example of passenger seating in an intercity (IC) train car

position in the car are registered by the observer. Then, the observer indicates to the norm breaker inconspicuously that he has finished recording the situation data and the norm breaker can start the intervention. Then, the norm breaker starts playing the song 'Robot Rock' by Daft Punk on his mobile phone at an annoying volume.[2] The intervention ends if a passenger in the car sanctions the norm breaker or when the song ends after 4 minutes and 48 seconds. The norm breaker stops the music only if another passenger asks him directly to do so; mere gestures of disapproval, exclamations not directed at the norm breaker, or not comprising a request to stop the music are not counted as sanctions. During the intervention, the norm breaker avoids eye contact with the other passengers in order not to make it too easy for them to inflict negative sanctions on him. After the intervention, the norm breaker and observer leave the car one after the other. Before the next intervention, the norm breaker helps the observer to complete the form with the other passengers' characteristics and reactions that the observer might have missed from his position during the intervention.

The second study differs from the first study in three important respects. First, all interventions were conducted on the direct line between Zurich and Basel, which takes 53 minutes without stops. Second, the norm breaker was always female. However, unlike in our first study, five different pairs of norm breaker and observer conducted the interventions. Third, we systematically varied whether the silence norm was violated in a silent-area car or in a non-silent-area car. It was only possible to vary the car type in intercity express (ICE) trains, in which silent-area cars are marked as such with conspicuous signs and non-silent-area cars lack such signs. Sometimes, taking the ICE was not possible. In these cases, interventions were conducted in the same IC trains as in study 1. In total, 59 interventions were conducted and data on 823 passengers were collected. Everything else was done in the same way as in study 1. Table 2 summarizes the most important aspects of the two studies.

Table 2. Summary of study designs

	Study 1	Study 2
Line	Zurich–Bern (58 min)	Zurich–Basel (53 min)
Train types	IC	IC and ICE
Car types	Non-silent (IC)	Non-silent (IC); non-silent and silent (ICE)
Norm breaker	Male	Female
Interventions (N_1)	31	59
Passengers (N_2)	204	823

Data

The data we have collected have multiple levels. The two levels that are relevant for our data analysis are the train car level, at which the $N_1 = 90$ interventions took place, and the (individual) passenger level, with a total of $N_2 = 1{,}027$ cases. The most important car-level variables that we recorded are the car type, the number of passengers in the car (including the observer but not the norm breaker), and the minutes left at the start of an intervention until the train reaches its destination. Recall that no interventions were started in the last 15 minutes of the train ride. The most important passenger-level variables that we recorded are the passengers' activities, their discernable gender and estimated age, and their location (i.e. seat) in the car. Based on each passenger's seat number and the seat number of the norm breaker (both the norm breaker's and the observer's seat were also recorded), each passenger's distance to the norm breaker (in terms of number of seats) was calculated. Table 3 lists the descriptive statistics of these variables. The data are available from the authors on request.

The main outcome variables that we recorded at the car level are whether or not the norm breaker was negatively sanctioned by another passenger and, if he or she was sanctioned, the time in seconds until the sanction occurred. At the passenger level, we also recorded the passenger who sanctioned the norm breaker. That is, we marked this passenger as 'sanctioner' on the data recording form (see Supplementary Figure SA1). In study 1, we also recorded passengers' reactions which did not qualify as sanctions. However, recording these reactions accurately was not always possible from the observer's location in the car. In order for the observer to be better able to focus on recording the contextual variables and passenger characteristics, we decided to discontinue recording passengers' other reactions in study 2. We therefore do not report results on passengers' other reactions in this article.

Results

Peer-sanctioning occurs in 50 per cent of our 90 interventions. The sanctioning rate is lower in study 1 (29 per cent) than in study 2 (61 per cent). This statistically significant difference ($\chi^2_{(1)} = 8.32$, $P = 0.004$) could be due to the fact that the norm breaker is male and female

Table 3. Descriptive statistics of car- and passenger-level variables

Variable	*N*	Missing	Mean	Median	*SD*	Minimum	Maximum
Car-level variables							
Car type	90	0	1.00				
Non-silent (ICE)	28	0	0.31				
Silent (ICE)	20	0	0.22				
Non-silent (IC)	42	0	0.47				
Number of passengers in car	90	0	12.39	10	7.10	2	30
Minutes to destination	90	0	34.82	35.5	12.41	13	57
Passenger-level variables							
Passenger's activity	1,027	0	1.00				
Wears earphones	71	0	0.07				
Talks to others or phone	251	0	0.24				
Reads or works	386	0	0.38				
Dozes, eats, looks window	219	0	0.21				
Other/unknown	100	0	0.10				
Is female	1,024	3	0.54				
Estimated age	1,024	3	39.55	35	16.80	5	75
Distance to norm breaker	1,027	0	13.57	12	9.24	1	46

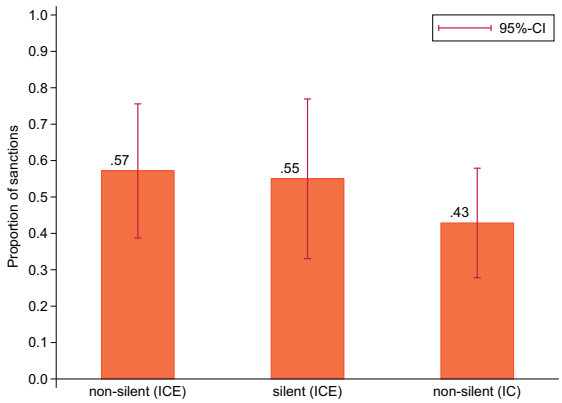

Figure 2. Proportion of sanctions across car types

in studies 1 and 2, respectively. However, since it is not only the norm breaker's gender that differs across the two studies (see Table 1), alternative explanations cannot be ruled out. These statistics should therefore be regarded as merely descriptive.

Surprisingly, as is apparent from Figure 2, the sanctioning rate does not differ significantly across car types ($\chi^2_{(2)} = 1.63$, $P = 0.443$). The silence norm is enforced at a slightly higher rate in ICE trains in non-silent-area cars (57 per cent) than in silent-area cars (55 per cent), and norm enforcement is lowest in IC trains in non-silent-area cars (43 per cent). The latter rate, although not significantly different from the other two, might be lower because most interventions in IC trains were conducted in study 1, where the norm breaker was male. In any case, this evidence supports neither H4 nor H4a; the two counteracting mechanisms hypothesized under H4 and H4a might be at work at the same time or not at all. We will come back to this point in the Discussion section.

We now turn to multiple regression analyses to further test our hypotheses (Table 4). We continue with testing our car-level hypotheses (H1, H3, H4, and H4a); in the second part of this section, we will test our passenger-level hypotheses (H2-1 through H2-3). Apart from the variables measuring and operationalizing the concepts in our hypotheses, all models in Table 4 also account for time-constant unobserved (and observed) differences across the two studies.

The first model in Table 4 is a logistic regression of the probability that the silence norm will be enforced. Accounting for other factors, the evidence of no difference in enforcement rates between silent- and non-silent-area cars in ICE trains does not change; H4 and H4a remain unsupported. Moreover, we do not find support for the diffusion of responsibility effect hypothesized under H1. In fact, the statistically significant coefficient estimate suggests that the more other passengers are in a car, the more likely is the silence norm enforced. Finally, we find support for H3. The closer the train is to its destination, the less likely it is that the silence norm will be enforced. However, using models for binary outcome variables, such as logit, does not entirely live up to the process that generated our data. In particular, these models do not account for the fact that our dependent variable

Table 4. Regression models

	Sanction occurred (0/1)	Time to sanction (cens. at 288")	Passenger-sanctioned norm breaker (0/1)	
Explanatory variables	**Logit 1**	**Cox PHM**	**Logit 2**	**FE logit**
H1: number of passengers in car	0.141**	0.071**	−0.033	
	(0.047)	(0.023)	(0.024)	
H3: minutes to destination	0.050*	0.021^{+}	0.026**	
	(0.021)	(0.012)	(0.009)	
H4, H4a: car type (non-silent ICE car is ref. cat.)				
Silent (ICE)	−0.260	−0.092	0.047	
	(0.617)	(0.371)	(0.245)	
Non-silent (IC)	2.424*	1.207**	0.498	
	(0.930)	(0.424)	(0.328)	
H2-1: passenger's distance to the norm breaker			−0.137***	−0.128***
			(0.030)	(0.029)
H2-2: passenger's activity (reads or works is ref. cat.)				
Wears earphones			−1.154	−1.915^{+}
			(0.706)	(1.070)
Talks with others or on the phone			−0.943*	−0.812
			(0.400)	(0.501)
Dozes, eats, looks out of window, etc.			−0.953*	−0.973^{+}
			(0.476)	(0.510)
Other/unknown			−0.445	−0.527
			(0.572)	(0.700)
H2-3: passenger's gender (male is ref. cat.)				
Female			−0.127	−0.378
			(0.328)	(0.343)
Control (study 1 is ref. cat.)				
Study 2	2.985**	1.682***	1.369***	Car FE
	(0.960)	(0.456)	(0.376)	
Constant	−6.516***		−2.845***	
	(1.618)		(0.792)	
N_1	90	90	90	45
N_2			1024	619
pseudo R^2	0.21	0.06	0.13	0.16
$\chi^2_{(df)}$	17.64 (5)**	24.21 (5)***	55.84 (11)***	35.82 (6)***

Notes: The table lists coefficient estimates with standard errors in parentheses ($^{+}P < 0.1$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, for two-sided tests). The first two models are based on car-level data only and estimate the probability (Logit 1) and the speed (Cox PHM) with which a sanction occurs, conditional on car-level characteristics. The two models are estimated with heteroscedasticity robust standard errors. Note that ICE car type is only varied in study 2, in which the norm breaker's gender is kept constant at 'female'. The last two models are based on passenger-level data and estimate the probability of a passenger to enforce the silence norm. The third model (Logit 2) accounts for car-level clustering and the fourth model (FE Logit) accounts for car FE. See the Supplementary Material for further regression model estimations.

not only tells us whether a sanction occurred or not, but also when it occurred, if it did. If the norm breaker was negatively sanctioned, the sanction occurred after 124 seconds on average (median = 120 s; SD = 87.1). What is more, our dependent variable is censored at 288 seconds—the length of the song played to provoke negative sanctions. Our second model accounts for these characteristics of our dependent variable.

The second model in Table 4 is a proportional-hazard model (PHM) for continuous time-to-event data, aka Cox regression (e.g. Hosmer, Lemeshow and May, 2008). Here too we find a higher number of passengers to be positively related with an earlier occurrence of sanctions, no difference between silent- and non-silent-area cars in ICE trains, and a decreasing propensity to enforce the silence norm over time. Although the coefficient estimate of the 'minutes to destination' variable is

now only significant at the 10 per cent level ($P = 0.074$), we can conclude that our results thus far are robust across different model specifications.[3]

Before we turn to testing our passenger-level hypotheses, we will first discuss the results of our third model with regard to the three car-level variables. Model 3 in Table 4 is a logistic regression of the probability that the silence norm will be enforced by a particular passenger. Note first that the coefficient estimate for the number of passengers, although statistically insignificant, is now negative. This does not mean that once we control for possible heterogeneity among the passengers in a car, we might be able to identify a diffusion of responsibility effect (H1). If anything, the negative coefficient estimate is an artefact. Since there can be only one sanctioner (recall that an intervention ends as soon as a negative sanction occurs), larger group sizes decrease the relative frequency of a single sanctioner. This is *not* the same as the diffusion of responsibility effect, which we derived from the symmetric VOD and tested using car-level models (see above). Moreover, model 4 does not give new evidence regarding H4 and H4a, and it confirms that it is indeed the individual passengers' propensity to enforce the silence norm that decreases with the train approaching its destination (H3).

As mentioned earlier, our train setting is more likely to resemble an asymmetric than a symmetric VOD. Asymmetry implies that passengers differ in the costs of and benefits from enforcing the silence norm. H2-1 through H2-3 are based on this assumption and predict that passengers with a higher net benefit from enforcing the silence norm will be more likely to sanction the norm breaker. We find good evidence in support of two of the three hypotheses. First, passengers who sit closer to the norm breaker are significantly more likely to enforce the silence norm (H2-1). Secondly, passengers who are engaged in activities that require silence are significantly more likely to enforce the silence norm (H2-2). That is, passengers who read or work are more likely to enforce the silence norm than passengers who wear earphones (although $P = 0.102$), who talk with others or on the phone, or who doze, eat or look out of the window. Finally, we find no support for H2-3. Although female passengers tend to be less likely to sanction the norm breaker, the negative coefficient estimate is statistically insignificant.[4]

In the last model in Table 4, we also account for car fixed effects (FE). Since the estimation of the FE logit is entirely based on the within-car variation in outcome and explanatory variables, we lose 405 cases and all 'car-constant' explanatory variables are dropped from the analysis (Halaby, 2004; Snijders and Bosker, 2012). However, estimating the FE logit has, in our case, three important advantages: First, observable and unobservable confounders which are constant within car are cancelled out in the estimation process as well. Most notably, passengers' self-selection into silent and non-silent cars becomes less of an issue for the test of our passenger-level hypotheses because car type is car-constant. Second, the FE logit is equivalent to the conditional logit (Long, 1997), which is commonly used to analyse choice behaviour in situations in which actors have several options from which they can choose only one (e.g. means of transport to go to work). In such decision situations, the choice options are not independent because choosing one option implies discarding all the other options. This is equivalent to the sanctioning situation we analyse here; a passenger who sanctions the norm breaker deprives all other passengers of the possibility to sanction the norm breaker. Third, since the estimation of the FE logit is entirely based on within-car variation of explanatory variables, and therefore on relative rather than absolute passenger characteristics, it is the best model to test our passenger-level hypotheses, which are derived from the asymmetric VOD.

We find our previous results hardly affected. First, there is still a substantial and statistically significant negative effect of a passenger's distance to the source of noise on the passenger's propensity to enforce the silence norm. Secondly, passengers who read or work are still more likely to sanction the norm breaker than passengers wearing earphones, talking with others or on the phone (although $P = 0.105$), or dozing, eating or looking out of the window. Finally, although now substantially stronger, the lower propensity of female passengers to sanction norm breakers remains statistically insignificant.

Discussion

Our article contributes to the research on how peer-sanctioning promotes cooperation and enforces social norms in two important respects. First, we argue that most research on peer-sanctioning has by and large neglected the strategic nature of the sanctioning situation. Many everyday norm violations produce a demand for negative sanctions in a group of people, which can be satisfied by one person alone. In these cases, a coordination problem can arise with regard to which group member should impose a sanction on the norm breaker. This second-order free-rider problem can be formally described by a step-level public good game such as the VOD (Diekmann, 1985, 1993; Raihani and Bshary, 2011). Previous research using computerized laboratory

experiments has shown that conceptualizing the second-order free-rider problem as a VOD can lead to new insights into the mechanisms of second-order public good provision. In particular, these studies have demonstrated that sanctioning cost heterogeneity has a major impact on the efficiency and effectiveness with which peer-sanctioning promotes (first-order) cooperation (Przepiorka and Diekmann, 2013; Diekmann and Przepiorka, 2015).

Our second major contribution to the literature lies in our extending the validity of these previous findings by showing that similar results can be obtained in domains of social life which resemble the set-ups created in the experimental laboratory. We conduct a quasi-experiment to investigate the enforcement of the silence norm by groups of passengers sitting in open-plan train cars. In 90 instances, we let a confederate play annoying music on his/her mobile phone and measure the time until a sanction occurs. Playing annoying music in an open-plan train car produces negative externalities and creates a second-order cooperation problem that resembles the VOD. In our quasi-experiment, the norm breaker was sanctioned and the silence norm was enforced in 45 out of 90 instances.

Furthermore, our findings support hypotheses derived from the asymmetric VOD (H2-1 through H2-3), in which the interacting parties are assumed to differ in their net benefits from producing the second-order public good. We find that passengers with lower costs of and/or higher benefits from sanctioning the norm breaker are more likely to do so. At the same time, our evidence clearly speaks against the diffusion of responsibility (H1)—the hypothesis that the sanctioning rate will decrease with group size. We find—to the contrary—that the more passengers are in a car, the more likely it is that the silence norm will be enforced.

A possible explanation for this unexpected result could be that in a larger group, passengers perceive it as more likely that someone will help them if the norm breaker reacts aggressively to their sanction. Hence, passengers' expected sanctioning costs are lower the more other passengers are present and therefore they are more likely to sanction the norm breaker (see Weesie and Franzen, 1998 on cost sharing in the VOD). Another explanation could be that some passengers want to maintain a good self-image by publicly enforcing social norms and the extent to which their self-image is boosted depends on the number of observers approving their behaviour. We consider explanations based on reputational incentives as less plausible because most passengers are unlikely to meet again in the future (see e.g. Raihani and Bshary, 2015; Przepiorka and Liebe 2016).

More obvious expectations find no support in our data. First, we do not find a significantly higher propensity of male passengers to enforce the silence norm. Previous studies of peer-sanctioning report mixed results concerning gender (see the literature review in the Supplementary Material). We suggest that future studies on norm enforcement try to disentangle gender effects from correlated factors such as body height, by systematically varying both, the norm breakers' gender and factors commonly related to gender. Second, norm enforcement rates do not differ across train cars with and without a sign making a silence norm being in effect explicit. On the one hand, we expected that passengers sitting in a silent-area car would feel more entitled to sanction the norm breaker and therefore would do so more than passengers sitting in a non-silent-area car (H4). On the other hand, breaking a norm that is made explicit by clearly visible signs may induce fear of retaliation in passengers, as the norm breaker will be more likely perceived as someone seeking a quarrel. Fear of retaliation may reduce the enforcement rate in silent-area cars as compared to non-silent-area cars (H4a). Since the two mechanisms hypothesized under H4 and H4a work in opposite directions, our results are inconclusive inasmuch as they leave open whether the two mechanisms neutralize each other or are not at work at all.

This is also where the limitations of our study become apparent. As the title of this article reveals already, ours is not an experiment in which subjects are randomly assigned to experimental conditions. However, random assignment would have greatly facilitated the exclusion of potential confounders. For example, some passengers choose silent-area cars exactly because they embrace a general reluctance to reprimand others for their rude behaviour, possibly wrongly believing that others sitting in silent-area cars with them will do it on their behalf. Future studies aiming at disentangling the two mechanisms hypothesized under H4 and H4a will need to find (or create) a setting, in which the norm breaker can be negatively sanctioned but cannot respond to the sanction. Only then will it be ruled out that fear of retaliation might curb the entitlement to negatively sanction a norm breaker.

Despite its limitations, the study at hand shows that it can be fruitful to model social interactions in general, and situations of social norm enforcement in particular, in game theoretic terms. Individuals' considerations to sanction norm breakers do not evade strategic thinking, which accounts for the costs and benefits of their and other bystanders' potential actions. Also in line with this idea are our findings in support of H3, suggesting that passengers indeed acknowledge the decreasing benefit of

enforcing the silence norm as the train approaches its destination. Confirming one of Olson's (1971 [1965]) early intuitions, our study shows that as long as the benefits outweigh the costs of producing the (second-order) public good, it can be in an actor's self-interest to produce it for the entire group by him or herself. Game theoretic models can help to discern situations with positive and negative net benefits from norm enforcement. For example, someone littering in a train station may not produce enough negative externalities to make it worthwhile for an observer to engage in peer-sanctioning. This could explain the low enforcement rate of the 'do not litter in public places' norm reported by Balafoutas and Nikiforakis (2012) (see e.g. Berger and Hevenstone 2016).

Notes

1 Depending on the research field, the informal enforcement of social norms is called negative sanctioning (sociology), peer-punishment (economics), or social control (social psychology), although definitions of informal norm enforcement can vary even within disciplines. Here, we use the terms (informal) norm enforcement and (negative) peer-sanctioning interchangeably.

2 The song can be listened to here: https://www.youtube.com/watch?v=HdeYwObD-j4 (retrieved 1 February 2016). Within each of the two studies, the volume at which the song is played is kept constant across interventions.

3 We performed further robustness checks and diagnostic tests on our car-level models. All these checks and tests bolster the conclusions we draw based on our car-level model estimations. We describe these additional analyses in full detail in the Supplementary Material sections S2.1 and S2.2.

4 We also estimated a model interacting norm breaker's gender with passenger's gender, and we estimated a mixed-effects model with random intercepts at the car level. These additional analyses did not produce different results or new insights (see the Supplementary Material for details).

Acknowledgements

Earlier versions of this article were presented at the Research Colloquium of the Institute of Sociology at University of Bern, the Institutions, Inequalities, and Life Courses Seminar of the Amsterdam Institute for Social Science Research at University of Amsterdam, the International Conference on Social Norms and Institutions in Ascona, and the 8th Maastrich Behavioral and Experimental Economics Symposium at Maastricht University. The authors would like to thank the organizers and participants of these events for their helpful comments. They would also like to thank four anonymous reviewers for their helpful comments and suggestions.

Funding

Chair of Sociology, ETH Zurich.

Supplementary Data

Supplementary data are available at *ESR* online.

References

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, **80**, 1095–1111.

Balafoutas, L. and Nikiforakis, N. (2012). Norm enforcement in the city: a natural field experiment. *European Economic Review*, **56**, 1773–1785.

Balafoutas, L., Nikiforakis, N. and Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences of the USA*, **111**, 15924–15927.

Berger, J. and Hevenstone, D. (2016). Norm enforcement in the city revisited: an international field experiment of altruistic punishment, norm maintenance, and broken window. *Rationality and Society*, forthcoming. http://dx.doi.org/10.1177/1043463116634035.

Bliss, C. and Nalebuff, B. (1984). Dragon-slaying and ballroom dancing: the private supply of a public good. *Journal of Public Economics*, **25**, 1–12.

Brauer, M. and Chekroun, P. (2005). The relationship between perceived violation of social norms and social control: situational factors influencing the reaction to deviance. *Journal of Applied Social Psychology*, **35**, 1519–1539.

Chekroun, P. and Brauer, M. (2002). The bystander effect and social control behavior: the effect of the presence of others on people's reactions to norm violations. *European Journal of Social Psychology*, **32**, 853–867.

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, MA: The Belknap Press of Harvard University Press.

Darley, J. M. and Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology*, **8**, 377–383.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution*, **29**, 605–610.

Diekmann, A. (1993). Cooperation in an asymmetric volunteer's dilemma game: theory and experimental evidence. *International Journal of Game Theory*, **22**, 75–85.

Diekmann, A. *et al*. (1996). Social status and aggression: a field study analyzed by survival analysis. *Journal of Social Psychology*, **136**, 761–768.

Diekmann, A. and Przepiorka, W. (2015). Punitive preferences, monetary incentives and tacit coordination in the punishment of defectors promote cooperation in humans. *Scientific Reports*, **5**, 10321.

Dreber, A. *et al* (2008). Winners don't punish. *Nature*, **452**, 348–351.

Fehr, E., Fischbacher, U. and Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, **13**, 1–25.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, **415**, 137–140.

Feinberg, M., Willer, R. and Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science*, **25**, 656–664.

Gächter, S., Renner, E. and Sefton, M. (2008). The long-run benefits of punishment. *Science*, **322**, 1510.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, **206**, 169–179.

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, **35**, 1–15.

Halaby, C. N. (2004). Panel models in sociological research: theory into practice. *Annual Review of Sociology*, **30**, 507–544.

He, J.-Z., Wang, R.-W. and Li, Y-T. (2014). Evolutionary stability in the asymmetric volunteer's dilemma. *PLoS ONE*, **9**, e103931.

Heckathorn, D. D. (1989). Collective action and the second-order free-rider problem. *Rationality and Society*, **1**, 78–100.

Herrmann, B., Thöni, C. and Gächter, S. (2008). Antisocial punishment across societies. *Science*, **319**, 1362–1367.

Horne, C. (2001). Sociological perspectives on the emergence of social norms. In Hechter, M. and Opp, K.-D. (Eds.). *Social Norms*. New York: Russell Sage Foundation, pp. 3–34.

Horne, C. (2009). *The Rewards of Punishment: A Relational Theory of Norm Enforcement*. Stanford, CA: Stanford University Press.

Hosmer, D. W., Lemeshow, S. and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Hoboken, NJ: John Wiley & Sons.

Krasnow, M. M. *et al*. (2012). What are punishment and reputation for? *PLoS ONE*, **7**, e45662.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, **21**, 53–74.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Lueptow, L. B., Garovich-Szabo, L. and Lueptow, M. B. (2001). Social change and the persistence of sex typing: 1974–1997. *Social Forces*, **80**, 1–36.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: can we really govern ourselves? *Journal of Public Economics*, **92**, 91–112.

Olson, M. (1971 [1965]). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, **14**,137–158.

Ostrom, E., Walker, J. and Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *American Political Science Review*, **86**, 404–417.

Palfrey, T. R. and Rosenthal, H. (1984). Participation and the provision of discrete public goods: a strategic analysis. *Journal of Public Economics*, **24**, 171–193.

Przepiorka, W. and Diekmann, A. (2013). Individual heterogeneity and costly punishment: a volunteer's dilemma. *Proceedings of the Royal Society B*, **280**, 20130247.

Przepiorka, W. and Liebe, U. (2016). Generosity is a sign of trustworthiness—the punishment of selfishness is not. *Evolution and Human Behavior*, forthcoming. http://dx.doi.org/10.1016/j.evolhumbehav.2015.12.003.

Raihani, N. J. and Bshary, R. (2011). The evolution of punishment in n-player public goods games: a volunteer's dilemma. *Evolution*, **65**, 2725–2728.

Raihani, N. J. and Bshary, R. (2015). The reputation of punishers. *Trends in Ecology and Evolution*, **30**, 98–103.

Ridgeway, C. L. and Correll, S. J. (2004). Unpacking the gender system: a theoretical perspective on gender beliefs and social relations. *Gender and Society*, **18**, 510–531.

Roberts, G. (2013). When punishment pays. *PLoS ONE*, **8**, e57378.

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Snijders, T. A. B. and Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Los Angeles, CA: Sage.

Ullmann-Margalit, E. (1977). *The emergence of norms*. Oxford: Oxford University Press.

Wolbring, T., Bozoyan, C. and Langner, D. (2013). "Links gehen, rechts stehen!" Ein Feldexperiment zur Durchsetzung informeller Normen auf Rolltreppen ("Walk Left, Stand Right!" A Field Experiment on the Enforcement of Informal Norms on Escalators). *Zeitschrift für Soziologie*, **42**, 239–258.

Weesie, J. (1993). Asymmetry and timing in the volunteer's dilemma. *Journal of Conflict Resolution*, **37**, 569–590.

Weesie, J. and Franzen, A. (1998). Cost sharing in a volunteer's dilemma. *Journal of Conflict Resolution*, **42**, 600–618.

West, S. A., El Mouden, C. and Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, **32**, 231–262.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, **51**, 110–116.

Wojtek Przepiorka is an Assistant Professor at the Department of Sociology in Utrecht. Before moving to the Netherlands, he held a Research Fellow position at Nuffield College and the Department of Sociology in Oxford. His research interests are in organizational behaviour, analytical and economic sociology, game theory, and quantitative methodology, in particular experimental methods. His most recent publications include 'Punitive Preferences, Monetary Incentives and Tacit Coordination in the Punishment of Defectors Promote Cooperation in Humans' (*Scientific Reports*, 5, with A. Diekmann) and 'Responsibility Attribution for Collective Decision Makers' (*American Journal of Political Science*, 59, with R. Duch and R. Stevenson).

Joël Berger is a visiting Research Fellow at the Department of Sociology in Groningen. Joël studied sociology and education in Bern, and he gained his PhD at ETH Zurich. His main fields of interest lie in competition and social inequalities, cooperation, experimental sociology, and applied game theory. Recent publications include: 'The Logic of Relative Frustration: Boudon's Competition Model and Experimental Evidence (*European Sociological Review*, 31, with A. Diekmann); 'Eye Spots Do Not Increase Altruism in Children' (*Evolution and Human Behaviour*, 36, with S. Vogt, C. Efferson and E. Fehr).