# Norm enforcement in the city revisited: An international field experiment of altruistic punishment, norm maintenance, and broken windows

*Joël Berger (ETH Zurich) & Debra Hevenstone (University of Bern/University of Zurich)*

**Abstract**

In laboratory experiments people are willing to sanction norms at a cost – a behavioral tendency called altruistic punishment. However, the degree to which these findings can be generalized to real-world interactions is still debated. Only a small number of field experiments have been conducted and initial results suggest that punishment is less frequent outside of the lab. This study replicates one of the first field experiments on altruistic punishment and builds ties to research on norm compliance and the broken windows theory. The original study addressed the enforcement of the anti-littering norm in Athens. We replicate this study in Bern, Zurich, and New York City. As an extension, we investigate how the experimental context (clean vs. littered) impacts social norm enforcement. As a second extension, we investigate how opportunity structure impacts the maintenance of the anti-littering norm. Findings indicate that norms are universally enforced, although significantly less than in the standard laboratory experiment, and that enforcement is significantly more common in Switzerland than in New York. Moreover, individuals prefer more subtle forms of enforcement to direct punishment. We also find that enforcement is less frequent in littered than in clean contexts, suggesting that broken windows might not only foster deviant behavior but also weaken informal social control. Finally, we find that opportunity structure can encourage people to maintain norms, as indicated by the fact that people are more likely to voluntarily pick up litter when it is closer to a trash bin.

**Keywords**

Broken windows, field experiment, norm enforcement, punishment, social control

Working Paper, January 2016.

An edited version of this paper has been accepted for publication in *Rationality and Society*. Online first: March 2016, DOI: 10.1177/1043463116634035.

## Introduction

Social norms regulate interactions between individuals in all known societies (Fehr and Fischbacher, 2004; Hechter and Opp, 2001). In many instances social norms constrain individual behaviors that have negative externalities (Ullmann-Margalit, 1977) and as such, help stabilize social order (Hardin 2013).

To ensure that social norms are followed, they must be enforced (Oliver 1980). Sanctions can be positive, i.e. reward norm compliance, or negative, i.e. punish deviance. Our study focuses on *negative sanctions,* as does most of the experimental literature on norm enforcement in anonymous one-shot interactions.[1] As enforcement is costly (e.g. in terms of effort and/or time) there is an incentive to free ride, leaving enforcement to others (Heckathorn, 1989; Yamagishi, 1986). Sociological research has identified mechanisms that overcome this "second-order free rider problem." For instance, enforcers might indirectly profit via reputation, status, or social capital or could even be directly rewarded by other group members (Axelrod, 1986; Horne, 2004, 2007; Willer, 2009) although not all studies find that enforcers are rewarded (Balafoutas, Nikiforakis, and Rockenbach, 2014). However, mechanisms based on gains in reputation, social status or social reward can only work in repeated, non-anonymous interactions.[2] It is still debated whether and, if so, how norms are enforced in the anonymous one-shot encounters that are common in contemporary societies (Guala 2012).

Evidence from laboratory experiments indicates that people are willing to pay to punish individuals who violate social norms even in anonymous one-shot interactions – a behavioral tendency called *altruistic* (or *costly*) *punishment* (Fehr and Gächter, 2002). Based on this observation, it has been suggested that, "the problem of second-order public goods can be solved if enough humans have a tendency for altruistic punishment" (Fehr and Gächter, 2002: 137).

Further laboratory studies have challenged the idea that altruistic punishment is the main mechanism sustaining social norms in the real world. In the lab, subjects are more reluctant to punish violators when the cost of sanctions is high (Anderson and Putterman, 2006; Carpenter, 2007; Egas and Riedl, 2008) or when retaliation is possible (Janssen and Bushman, 2008) – two key factors in real world-interactions. As such, Guala (2012) objects that it cannot be concluded from existing laboratory evidence that punishment is a common and efficient way of maintaining social norms. Other limitations associated with laboratory experiments include the self-selection of cooperative individuals into subject pools (Harrison and List, 2004) and the potential over eagerness of participants to fulfill experimenters' expectations (i.e. demand effects, see Orne 1962). Moreover, in quotidian life, people usually have more options than just passivity or punishment (Guala 2012). These issues (i.e. self-selection of cooperative individuals, demand effects, and restriction of the subjects' scope of action) can be solved with natural field experiments.[3] In a natural field experiment (from now on: "field experiment"), the "participants" are not aware that they are taking part in a study and are observed in natural quotidian social interactions, thus overcoming the aforementioned limitations.

To date, only a small number of field experiments on social norm enforcement have been conducted. Initial results suggest that in natural interactions, people are less willing to enforce norms than in the laboratory (Balafoutas and Nikiforakis, 2012), that people prefer less costly forms of norm enforcement to direct punishment (Balafoutas , Nikiforakis, and Rockenbach, 2014), and that people who feel personally affected by norm violations are more likely to sanction (Brauer and Chekroun, 2005; Przepiorka and Berger, forthcoming). Moreover, there seem to be large disparities in enforcement rates across locations: Balafoutas, Nikiforakis, and Rockenbach (2014) find a strikingly higher enforcement rate of the anti-littering norm in Cologne than in Athens (Balafoutas and Nikiforakis, 2012).

Understanding why and when norms are enforced is of interest not only to basic social science, but also to social policy. The enforcement of some norms, like littering, is a public good, and creating conditions encouraging norm enforcement are thus a potential public policy tool. It is for this reason that the New York City Parks and Recreation website states, with respect to the new ban on smoking in parks, "the new law will be enforced mostly be New Yorkers themselves. We expect that New Yorkers will ask people to follow the law and stop smoking."[4]

The contribution of this study to the literature is fivefold. First, we replicate the study by Balafoutas and Nikiforakis (2012) in Bern, Zurich and New York. We thereby intend to assess the prevalence of enforcement and to replicate the finding that the prevalence of norm enforcement varies strikingly across places.

Second, we extend their design, bringing together two strands of research: research on norm enforcement on the one hand and research on norm compliance and the broken windows theory on the other. The broken windows theory states that signs of disorder and petty crime (e.g. litter or graffiti) can engender further disorder, norm-breaking behavior and even serious crime, thereby generating a snowball effect (Kelling and Coles, 1996; Wilson and Kelling, 1982). One hypothesized mechanism is that people infer from signs of disorder that norms and legal rules are rarely enforced and thus there is a low risk of getting caught (Keuschnigg and Wolbring 2015). An alternative mechanism is that signs of disorder trigger a change in one's frame of mind, from a normative frame to a hedonic frame. When in a hedonic frame, people tend to pursue hedonic goals, i.e. their immediate well-being, rather than follow norms (Keizer, Lindenberg, and Steg 2008). For example, they would drop a plastic bottle rather than hold onto it until finding a trashcan. Most research on the broken windows theory is based on observational and quasi-experimental data and results are inconclusive (e.g. Eck and Maguire, 2000; Harcourt and Ludwig, 2006; Kelling and Sousa, 2001; Sampson and Raudenbush, 1999), while more recent experimental studies demonstrate that indeed, people are less likely to conform with social norms and legal rules in the presence of cues implying that others are not following social norms either (Cialdini and Kallgren, 1990; Keizer, Lindenber, and Steg, 2008; Keuschnigg and Wolbring, 2015). Here we test the related hypothesis that if there is disorder, people are not only less likely to adhere to norms, but they are also less likely to enforce them. This would be a third potential mechanism for the snowball effect hypothesized by broken windows theory.

Third, we extend the baseline to design to test whether the opportunity structure of an environment is relevant for the maintenance of social norms. Policy makers can influence

the opportunity structure of an environment to promote the population to maintain social norms, thus reducing the probability of snowballing disorder at time zero (see Hemenway 2013; Hevenstone 2015: chapter 6; and Posner and Rasmusen, 1999: 381).

Fourth, we report reactions to norm violations in a more detailed manner, going beyond the binary response of punishment versus passivity that is available to subjects in the typical laboratory experiment (Guala 2012).

Fifth, we compare the individual characteristics of enforcers to potential enforcers, testing whether certain types of individuals are more likely to enforce and maintain social norms.

## Experimental design

### *Operationalization and treatments*

We chose to investigate the *anti-littering norm*, in part because it was tested in a prior field experiment on norm enforcement (Balafoutas and Nikiforakis, 2012). In addition, enforcing the anti-littering norm requires a relatively high cost in relation to the benefit, since the externality of one dropped piece of litter is rather small—an important attribute when studying altruistic punishment. In contrast, an individual enforcing the queuing norm (Milgram et al. 1986; Schmitt, Dubé, and Leclerc, 1992) reaps a large direct benefit so enforcement cannot incontrovertibly be attributed to altruism (see Diekmann and Przepiorka, 2015).[5]

We distinguish between four reactions to norm violations. First, a *direct sanction* (i.e. verbally confronting the norm violator) is clearly altruistic punishment as it is costly in terms of both time and effort, entails psychological costs (Adams and Mullen, 2012), and engenders the possibility of retaliation.

A second reaction, the *subtle sanction*, is when rather than confronting the violator, the bystander throws an angry glance at the violator or talks to others about the violator. This is an expression of disapproval with a much lower risk of retaliation. However, our results with regard to subtle sanctions should be interpreted with caution. Expressions of disapproval can be ambiguous or fleeting, and we only recorded those signs of disapproval that we observed and that were clear.

A third potential reaction is that people can *pick up litter* (a plastic bottle) that the experimenters dropped. Unlike subtle sanctions, this behavior can be unambiguously identified, but its meaning can be ambiguous. If the violator observes the person picking up the bottle, it is an indirect reprimand. Alternatively, if only bystanders observe it, it is a form of norm reinforcement in the community. Picking up a dropped bottle also directly contributes to the first-order public good of a clean environment (thus confounding maintaining the norm with norm enforcement). Maintaining the norm also potentially reduces further norm violations by maintaining the social norm of a clean environment in the first place (i.e. preventing a broken windows effect).[6] In sum, picking up the bottle, depending on the individual motivation, can be a form of norm enforcement or a contribution to norm maintenance.

The forth possibility is *no reaction*. One might say that only direct sanctions are altruistic punishment in the sense of Fehr and Gächter (2002) while subtle sanctions and picking up the bottle are alternative, less costly, actions that can also strengthen the anti-littering norm.

For the *broken windows treatment,* the experiment was conducted in clean and littered settings. In the *no litter condition* we removed surrounding litter if necessary, while in the *litter condition*; we placed a bag of garbage[7] and several pieces of litter in the experiment setting (see Figure 1). We conducted the broken windows treatment in Bern and New York City.

**Figure 1**. No litter condition and litter condition

In Zurich, we skipped the broken windows treatment because the effects were clear. Instead, we manipulated the *opportunity structure* to encourage or discourage norm maintenance by conducting the experiment alternatively *close to or distant from a trashcan*. We suspected that people would be less willing to pick up the bottle when distant from a trashcan, as they would then have to walk 12 meters to dispose of the bottle or to keep the bottle in their hands until the train arrived (there are also trashcans inside trains).

*Subjects, places, and procedures*

Data was collected on two types of people: "targets" and "enforcers." The target was the first individual between ages 18 through 70 who stood at a predefined spot after the last train or bus left the experimental area. The norm violation was conducted directly in front of these "targets," with the expectation that they would enforce. However, despite the targets' proximity to the violation, other bystanders enforced almost as often as the targets. For both targets and enforcers we collected measures of individual characteristics (estimated age, gender, and whether the individual was a member of the majority (in) group). As age was a rough approximation based on appearance, four analyses we recoded estimated age into dummies for young adults (age 18 to 25), adults (26 to 55), and older adults (56-70). In-group members were defined as members of the ethnic or racial majority (white, presumably Swiss people, in Switzerland and white, presumably American people, in New York City). For each experiment we also counted the number of people in the 3 x 7 meter area in which the violation was easily visible, to measure the bystander density.

The first experiment was conducted in *Bern,* the capital of Switzerland (population 138,000).[8] The experiment was conducted at a tram stop in front of the main rail station. The stop, *Hirschengraben*, is used by approximately 40,000 passengers a day.[9] We selected a central crowded stop to maximize anonymity and population diversity, running the experiments near the stop's trashcan. The individual standing nearest to the trashcan was defined as the *target*. One of the experimenters was assigned to the role of the *norm violator* and the other to the role of *observer*. Roles were exchanged after every 5 trials. Both experimenters (female, white, 36 years old, 1.73 m tall and male, white, 30 years old, and 1.83 m tall) wore blue jeans and a black shirt.[10]

The observer, who coded the variables of interest, was discretely seated on a bench behind an obstacle (a bike stand), 7 meters from the trashcan. The violator walked towards the trashcan and threw an empty plastic bottle from about 2 meters away. The bottle, missing the trashcan, fell to the floor near the target, and the violator continued walking without picking it up. To bystanders, this looked as if the violator was too lazy to pick up the bottle after having missed the trashcan. The experiment was never conducted when a tram was at the stop, as the noise made it difficult to hear the dropping bottle. The next trial was not conducted until all bystanders from the prior experiment had left. The experiment was conducted on four working days in July 2013 between 1.30 pm and 7.30 pm with the intention of encompassing both non-rush hour and rush hour traffic. The first experimental condition was determined randomly, and then was switched every 2.5 hours.

We replicated the experiment in *New York City* on two working days in August 2013 under similar conditions as in Bern. We selected two platforms at the Union Square subway

station. The intention of selecting a central, but not *the* central station (e.g. Grand Central or Penn Station), was to generate a representative sample of the local population, with lower volumes of pedestrian traffic, closer to the Bern case.[11] However, even as a secondary station, Union Square is still frequented about 2.5 times more than *Hirschengraben*; in 2012, 108, 000 people per day used Union Square station.[12]

Results from the first two experiments suggested that both direct enforcement and subtle enforcement is considerably more frequent in Bern than in New York (5.0 versus 2.6% of the time and 22.6 versus 15.4%, respectively) and – insofar as our design is comparable to Balafoutas and Nikiforakis (2012) and to Balafoutas, Nikiforakis, and Rockenbach (2014) – direct enforcement in Switzerland is more frequent than in Athens and New York City but not as frequent as in Cologne.[13] Concerned that the difference between enforcement in Bern and New York City might be due to city size rather than country or culture, we decided to run a third experiment in Zurich, the largest city in Switzerland with about 400,000 inhabitants (over 1.9 million in the metropolitan area). A lower difference in norm enforcement between Zurich and Bern versus between Zurich and NYC, would suggest that the difference is related to nationality rather than city size. The third experiment was conducted on two underground platforms in Zurich's main train station, which about 400,000 passengers use daily.[14] The experiment was conducted on three working days in September and October 2013. We conducted 174 trials, split between the baseline experiment, and a new treatment varying the opportunity structure for norm maintenance (i.e. proximity to trashcan), as described above.[15]

## Results

*Strategy of analyses*

We first report descriptive statistics and then marginal effects (Figures 2 to 4), stemming from the multinomial logit models listed in Table A1 in the appendix. The dependent variable is "reaction" (direct sanction, subtle sanction, and picking up the bottle, as compared to no reaction).[16] The experimental predictors of interest are "litter" ("no litter" as a reference category), "Bern" and "New York" (with "Zurich" as a reference category), "remote distance to trashcan" (with "close to trashcan" as a reference category). As control variables we added density, (the number of people present in a predefined area of 3 x 7 meters surrounding the spot of experimentation), density squared (to account for nonlinearity) and a dummy "rush hour" ("no rush hour" as a reference category). Our main model focuses any first reaction, by target or other bystanders. Additionally, we present results from a model analyzing secondary reactions, occurring only after a first person already had reacted, and results from a model focusing on only "target" reactions, as individual characteristics cannot be controlled for in the model predicting any reaction, as the individual characteristics of all bystanders was not collected.

*Descriptive results*

Overall, in 32.0% of all 488 trials did at least one person react, either directly (direct sanction 4.5%), with an expression of disapproval (subtle sanction, 14.3% probably underestimated due to conservative coding approach) or, with picking up the bottle (norm maintenance, 10.3%). In 2.87% of all trials, a person had a joint reaction, generally combining a direct sanction with picking up the bottle or a subtle reaction and picking up

the bottle. There was never more than one double reaction in a given trial. Slightly more than half of all reactions (56.67%) were target-reactions with the rest coming from other bystanders. We also recorded secondary reactions: in approximately 6.4% of all trials a second person joined in after a first person had already taken action.

*Place*

Considering the location of the experiment (Figure 2), there are substantial differences between Bern and Zurich compared to New York City. Based on the multinomial logit model predicting any first reaction (see Table A1 in the appendix), we expect in the baseline condition (without litter and close to the trash bin) that direct norm enforcement would occur about 10% of the time in Bern, 9% of the time in Zurich, and just 4% in NYC (marginal effects, holding density and rush hours at the mean. New York City differs from Bern with a $p$ value of .051 or from Zurich $p$ = .10, while Bern and Zurich do not significantly differ). Picking up the bottle is more frequent in Switzerland, too. Predicting the chances of picking up the bottle, again under the baseline experimental conditions and mean density and rush hour, in Zurich the bottle would be picked up a full 27% of the time compared to 13% in Bern and just 9% in NYC. (Bern is not statistically different from Zurich or NYC, but NYC is statistically significant from Zurich with $p$ = .008.). Marginal effects for subtle sanctions are higher in Bern (24.5% of all trials) than in NYC (18.20%, $p$ = .081), while Zurich is incomparable because we curtailed data collection. In sum, all three reactions, direct sanctions, subtle sanctions, and picking up the bottle occur more frequently in Switzerland than in NYC.

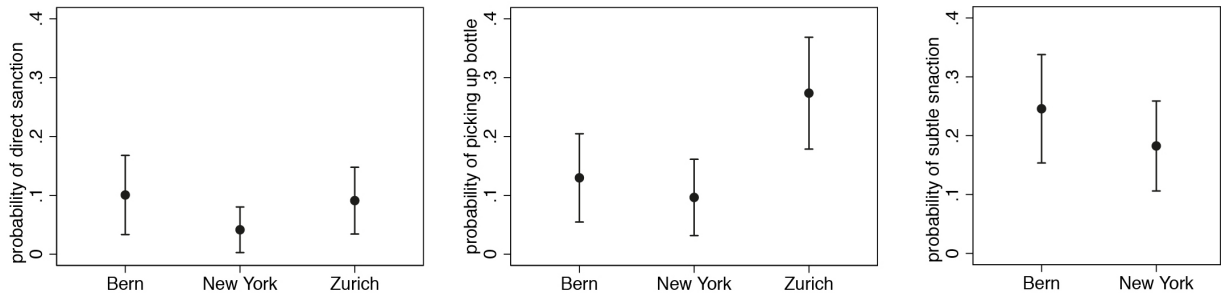


**Figure 2.** Predicted probability of direct sanction, bottle picking, and subtle sanction by city

*Disorder*

Disorder considerably impacts norm enforcement (Figure 3). Marginal effects (again holding environmental factors at the mean), suggest the violator is directly confronted just 2.7% of the time under the littered condition compared to 7.7% of the time with no litter ($p$ = .024). The effect is even more pronounced for picking up the bottle, which drops from 16.2% in the no litter treatment to 1.41% in the litter treatment ($p$ = .002) and somewhat weaker for subtle sanctions, which drop from 14.5% to 10.4% ($p$ = .030).

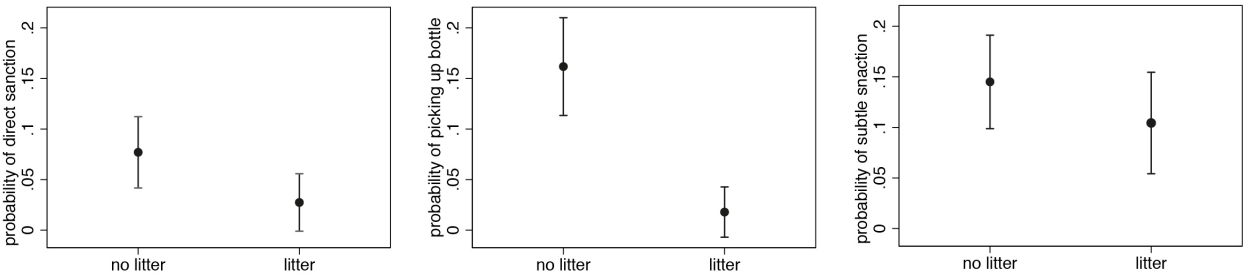


**Figure 3.** Predicted probability of direct sanction, bottle picking, and subtle sanction by disorder-treatment

*Opportunity structure*

We conducted trials close to and far from a trashcan (garbage bin) with the intent of modifying the opportunity structure to maintain the no littering norm. If the experiment is conducted close to the bin, bystanders can pick up the bottle and dispose of it immediately. In contrast, far from the bin (12-meters away), they must either walk to the trashcan or keep the bottle until the next train arrives. In other words, the costs of maintaining the norm are higher when there is no trashcan nearby. In contrast, this manipulation should not impact direct sanctions (altruistic punishment).

As can be seen from Figure 4, there is a marginal effect of 16.2% of all trials in which a bystander would pick up a bottle dropped close to a trashcan as compared to 3.6% when it is dropped far from the trashcan ($p$ = .002). The distance to the trashcan has no statistically significant effect on the prevalence of direct sanctions. This would suggest that picking up the bottle and direct sanctions are not substitutes.[17] We cannot say anything about the effect of the distance to the trashcan on subtle sanctions because we implemented the distance treatment in Zurich where subtle sanctions were not recorded.

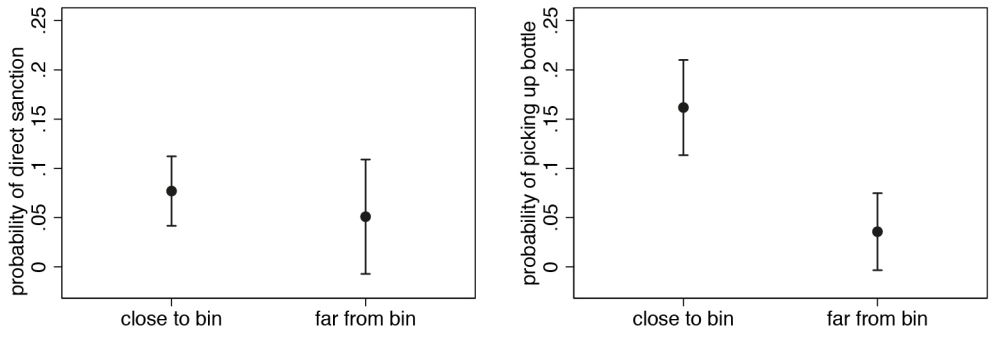


**Figure 4.** Predicted probability of direct sanction and bottle picking by distance from the trashcan

*Contextual variables*

We did not find a significant effect of the contextual variables density/density squared or rush hour with the exception of a small negative effect of rush hour on the probability of secondary reactions and an increasing and then stabilizing effect of density on direct sanctions.[18] The fact that density, at least initially, seems to increase the likelihood of a direct sanction, suggests that there is no diffusion of responsibility effect (see Chekroun and Brauer, 2002). While rush hour is not significant, it is consistently negative, suggesting perhaps a slight tendency towards fewer sanctions or less norm maintenance during rush hour, which makes sense since commuters generally rush to catch the train.

*Individual Characteristics*

There are two ways we can understand how individual characteristics are correlated with sanctioning. We had information on individual characteristics of both responding and non-responding targets, but not information on non-responding bystanders. (Characteristics include age, in-group, and gender.) Given this data limitation, first, we report a multinomial logit predicting the likelihood that *targets* responded (see Table A1 in the appendix). However, excluding responding non-targets, the sample size of enforcers drops. As such, we complement this regression analysis with a second analysis in which we compare enforcers to the target group. The target group can be considered a random sample of people at the bus stop, since we chose targets based on their proximity to the trashcan. As such, the average traits of bystander enforcers can be compared to the targets, to understand how enforcers differ from the average population.

Results from the multinomial logit model suggest that out-group members are less likely to express subtle sanctions than in-group members (marginal effects of 5.3% vs. 14.4%; $p$ = 0.01). Coefficient sign suggests out-group members are also less likely to engage in direct sanctions, though the effect is insignificant. Older adults are significantly more likely to engage in direct sanctioning with a marginal probability of 4.5% compared to 1.4% among young adults and .6% among mid-aged adults ($p$ = .019). With respect to gender, there are no significant effects. Results with respect to individual characteristics were generally weak because of the low numbers, which is why we also take a second strategy comparing all enforcers to the target group.

Table 1 illustrates the percent male, in-group, young adults, and older adults for all targets, first and later responders who engage in direct sanctions, first and later responders who pick up the bottle, and first and later responders who engage in subtle sanctions. Significance refers to a test of proportions compared with the target group. Results largely confirm those found using the multinomial logistic model. There were significantly more in-group individuals among those engaging in direct and subtle sanctions than in the random target group. Older adults were also more prevalent among those engaging in direct sanctions (though only secondary respondents), those picking up the bottle, and those using subtle sanctions. What differs from the regression model is that we also find that young adults are slightly more prevalent among those engaging in direct sanctions than those in the target group and that there are slightly more men among those engaging in subtle sanctions. Jointly, the two analyses strongly suggest that in-group members and older adults are more likely to take an active role in maintaining and enforcing a no-littering norm.

**Table 1.** Individual characteristics by group, compared to the randomly selected targets.

| | *Percent male* | *Percent in-group* | *Percent young adults* | *Percent older adults* |
|---|---|---|---|---|
| **All targets** | **54.10** | **75.41** | 23.41 | **7.60** |
| *direct sanctions* | | | | |
| first responders | 43.75 | 87.5** | 34.38* | 12.5 |
| later responders | 25.00 | 100.00 | 25 | 50.00*** |
| *picking up the bottle* | | | | |
| first responders | 53.70 | 67.19 | 26.69 | 20.31*** |
| later responders | 72.73 | 83.33 | 8.33 | 8.33 |
| *subtle sanctions* | | | | |
| first responders | 62.67* | 84.21** | 25.00 | 18.42*** |
| later responders | 42.11 | 72.73 | 27.27 | 22.73*** |

* .1 sig
** .05 sig
*** .01 sig

*Secondary reactions*

In 6.35% of all trials a second person joined in after a first person had taken action. For example, one person confronted the violator while the other picked up the bottle. The only significant predictor for secondary reaction was place, with a marginal probability of picking up the bottle of 10.8% in Zurich compared to 3.3% in Bern and 1.1% in NYC (see Table A1 in the appendix).[19]

## Discussion

We conducted field experiments on the enforcement of the anti-littering norm in Bern, New York City, and Zurich. Direct sanctions (confronting the violator), subtle sanctions (e.g. angry looks, shakes of the head, talking to other bystanders about the incidence) and norm maintenance (picking up the litter) were observed in all three cities, with higher rates of all three in Bern and Zurich than in New York. There are various potential reasons for the difference between countries. Our results might corroborate anecdotal evidence that there is generally more public norm enforcement in Switzerland than in the USA (Hevenstone, 2015). As further anecdotal evidence, in Bern we even had difficulties in maintaining the litter condition. Twice people wanted to remove the bag of garbage. This never happened in NYC.

A survey study by Brauer and Chaurand (2010) indicates that the level of social control in a society is negatively related to the degree of individualism as conceptualized by Hofstede (2001). As such, it might be the US's greater individualism that leads to less enforcement. However, results comparing Germany and Greece do not match up to differences in individualism. In Cologne the enforcement of the anti-littering norm was observed four times as often than in Athens while individualism in Germany is twice as high than in Greece (see Balafoutas, Nikiforakis, and Rockenbach, 2014 and Balafoutas and Nikiforakis, 2012, Hofstede 2001). It might also be the case that it is not norm enforcement in general that differs across places, but rather the importance of the specific norm. Perhaps anti-littering norms are stronger in Switzerland while anti-smoking norms are stronger in the

US. Another potential reason for the disparity in enforcement rates is fear of reprisal, which could be greater in New York City than in Zurich or Bern.[20] Another possibility is that the proportion of "altruistic punishers" differs between the countries. Finally, city size might partially account for these differences, although Zurich and Athens are comparable with respect to city size, while differing substantially in punishment rates. It is impossible to definitively determine the causes of inter-country differences with data on only four countries, and as such, is something to be addressed in future research.

A second finding is that in littered settings, bystanders are much less prone to either directly sanction a norm violator, to maintain the norm by picking up the bottle, or engage in subtle sanctions. This means that signs of disorder might not only weaken norm compliance (e.g. Keizer, Lindenberg, and Steg, 2008), but they also weaken social control and norm maintenance, which would be a second mechanism perpetuating the spreading of disorder.

A third finding is that the opportunity structure influences people's willingness to maintain social norms. When litter was dropped close to a trashcan, bystanders were much more likely to pick up the bottle than when it was dropped far from a trashcan. When norm-breaking behavior is diffuse and direct intervention by the authorities is not possible, policy makers should design environments such that the public maintains them. This could stop the spread of norm violations at an early stage. In the case of littering, a high density of trashcans could prevent the spreading of disorder (see Hemenway 2013; Hevenstone 2015: chapter 6; and Posner and Rasmusen, 1999: 381 for related examples).

As a fourth finding we observed subtle sanctions and norm maintenance are more frequent than direct sanctions (i.e. altruistic punishment). This concurs with Balafoutas, Nikiforakis, and Rockenbach (2014) who report that, when possible, people prefer to indirectly punish norm violators by withholding help to punishment. This finding supports Guala's (2012) call for field experiments, which should complement laboratory experiments in order to achieve a more fine-grained understanding of how norms are enforced in real-world conditions.

Additional findings include the fact that in-group members and older adults are more likely to enforce and maintain norms. Contextual factors (rush hour and density of bystanders in the area) had very weak effects. There were some small effects for increasing density initially increasing enforcement and then stabilizing, but we cannot be certain of this finding without further research (see Chekroun and Brauer, 2002 on diffusion of responsibility in norm enforcement). In addition, there were some indications that there might be somewhat less enforcement in rush hour.

In sum, altruistic punishment specifically (i.e. confronting transgressors), and norm enforcement generally, occur in natural anonymous one-shot interactions across societies. However norm enforcement (direct sanctions) and norm maintenance (e.g. picking up the bottle) are sensitive; they depend on place, context and opportunity structure; and the characteristics of the population at hand. Moreover, people prefer subtle sanctions and norm maintenance to direct confrontation. Future studies should further investigate the subtle forms of enforcement that are probably more widespread than actual punishment. In addition, a broader set of norms and international comparisons should be considered in

future studies to determine the causes of international differences. Moreover, the conditions under which norms are most likely to be enforced should be identified for each norm. This would be of interest both for social theory and for policy makers, who try to shape environments that foster the endogenous maintenance of social norms.

**Acknowledgements**

We thank Roger Berger, Benita Combet, Marc Höglinger and two anonymous reviewers for helpful comments on a previous version of this manuscript.

[1] For example, parents might praise their children for throwing a piece of paper in the garbage (positive) while a passerby might reprimand someone for littering (negative) (Coleman, 1990). For more information on positive sanctions in anonymous one-shot interactions see Berger (2011) and Diekmann, Jann, and Wehrli (2014).

[2] An enforcer cannot profit from his reputational gain when there are no future encounters with those individuals who observed the enforcer.

[3] Specifically, there are "lab experiments" conducted outside of the lab with non-student subjects. In these experiments standard laboratory conditions are generally applied: subjects know that they are taking part in an experiment, they know the rules of the game and payoffs, and understand that all decisions are anonymous. This is called a *framed field experiment* whereas our study is a *natural field experiment*, which is conducted in an everyday situation and where the subjects do not know that they are taking part in an experiment (Harrison and List, 2004; Levitt and List, 2007).

[4] http://www.nycgovparks.org/facility/rules/smoke-free, retrieved Dec 19, 2013.

[5] For example, Balafoutas and Nikiforakis (2012) also investigated the norm that on an escalator those not wishing to walk should stand on the right-hand side, allowing others to pass on the left. In this study, the enforcement rate was considerably higher (see Wolbring, Bozoyan, and Langner (2013) for a replication).

However, since there are individual incentives to enforce the escalators norm (an individual enforcing the norm might simply want to pass), it is questionable whether this is a suitable test of altruistic punishment.

[6] On the last day of experimentation in Zurich we asked norm-enforcers for their motives, using a standardized questionnaire. Individuals picking up the bottle reported both a desire to demonstrate their concern about littering (5 of 12 cases) and a desire to keep the station clean (7 cases).

[7] This is prohibited in Switzerland as in NYC.

[8] http://www.bern.ch/leben_in_bern/stadt/statistik/katost/01bev, retrieved July, 2014.

[9] Estimate by the public transit authority for 2012.

[10] One experimenter is American, the other Swiss. Because the experimenters did not speak when dropping a bottle and it is not possible to distinguish white US American people from Swiss people by just looking at them, nationality could not have impacted the results.

[11] By not choosing the most central station in NYC we increased the chance that our sample contained more New York City-dwellers and fewer tourists, who might be less familiar with social norms prevalent in the city. In Bern and Zurich by choosing commuting platforms in the central stations, we were able to use a location with levels of foot traffic similar to the non-central station in New York, and also with few tourists.

[12] http://www.mta.info/nyct/facts/ridership/ridership_sub.htm, retrieved December 28, 2013.

[13] This design is a slight deviation from the Athens and the Cologne experiments, where the target and the littering incident were the same. In these experiments, the authors managed to isolate the target and then drop a piece of litter in front of them. This procedure excludes the second-order free rider problem by design; when there are no bystanders, only the target individual can enforce the norm. After a few pilots, we found it impossible to isolate individuals and to prevent bystanders from becoming enforcers. This means that our procedure does not exclude the second-order free rider problem by design. Because bystanders could enforce the norm, there is the potential for a diffusion of responsibility—i.e., each individual has an incentive to leave the punishment of norm violators to the others. By including the second-order free rider problem, our design is closer to the standard public goods game laboratory experiment where experiments are conducted in groups of four.

[14] Estimate from Swiss Federal Railways.

[15] To be more specific, in the control condition we followed exactly the same procedure as in Bern and in NYC – the violator threw a plastic bottle at the trashcan and missed it. At a 12 m distance from the trashcan, the violator just dropped the bottle. As a bridge between both procedures, we also ran some experiments where the violator simply dropped the bottle beside the trashcan. Results for throwing the bottle at the trashcan and dropping it near the trashcan are identical. As such, we decided to pool the data for throwing and dropping near the trashcan.

[16] We also considered modeling joint reactions separately. However, there were only 10 cases of direct enforcement combined with picking up the bottle and just 4 cases of subtle sanctions and picking up the bottle. We could not model these joint reactions because incidence was so low. Instead we recoded joint reactions as the dominant reaction (i.e., direct enforcement with picking up the bottle was considered direct enforcement). The recoding of these 14 joint reactions into singular reactions did not change results.

[17] See also Balafoutas, Nikiforakis, and Rockenbach, 2014, who find that direct and indirect sanctions (i.e. withholding help) are substitutes.

[18] Weather could not have played a role as experiments were only conducted when it was not raining.

[19] One potential explanation for why people in Zurich picked up bottles more often than Bern might be that the Zurich station was underground (as was the NYC station), while the Bern station was outdoors. Potentially people might be more likely to enforce when they are "indoors."

[20] Bystanders in New York City might have a greater fear of reprisal because there is more violence in NYC than in Zurich or Bern.

**References**

Adams GS and Mullen E (2012) The social and psychological costs of punishing. *Behavioral and Brain Sciences* 35: 15-16.

Anderson CM and Putterman L (2006) Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior* 54: 1-24.

Axelrod R (1986) An evolutionary approach to norms. *The American Political Science Review* 80: 1095-111.

Balafoutas L and Nikiforakis N (2012) Norm enforcement in the city: A natural field experiment. *European Economic Review* 56: 1773-85.

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 United States of America* 111: 15924–15927.

Berger J (2011) Altruistic reciprocity. Empirical evidence from sequential dictator games and sequential taking Games among 11-year-old children. *Soziale Welt* 62: 165-81.

Brauer M and Chaurand N (2010) Descriptive norms, prescriptive norms, and social control: an intercultural comparison of people's reactions of uncivil behaviors. *European Journal of Social Psychology* 40: 490-499.

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: 1-22.

Carpenter JP (2007) The demand for punishment. *Journal of Economic Behavior & Organization* 62: 522-42.

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.

Cialdini RB, Reno RR and Kallgren CA (1990) A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58: 1015-26.

Coleman JS (1990) *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Diekmann A, and Przepiorka W (2015). Punitive preferences, monetary incentives and tacit coordination in the punishment of defectors promote cooperation in humans. *Scientific Reports* 15: DOI: 10.1038/srep10321.

Diekmann A, Jann B, Przepiorka W and Wehrli S. (2014) Reputation formation and the evolution of cooperation in anonymous online markets. *American Sociological* Review 79: 65-85.

Eck, JE, and Maguire, ER. (2000) Have changes in policing reduced violent crime? In: Blumstein A, Wallman, J (eds), *The Crime Drop in America*. Cambridge: Cambridge University Press.

Egas M, and Riedl A (2008) The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences* 275: 871-78.

Fehr E and Fischbacher U (2004) Third-party punishment and social norms. *Evolution and Human Behavior* 25: 63-78.

Fehr E and Gächter S (2002) Altruistic punishment in humans. *Nature* 415: 137-40.

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

Harcourt BE and Ludwig J (2006) Broken windows: New evidence from New York City and a five-city social experiment. University of Chicago Law Review 73: 271–320.

Hardin R (2013) The priority of social order. *Rationality and Society* 25: 407-421.

Harrison GW and List JA (2004) Field experiments. *Journal of Economic Literature* 42: 1009-55.

Hechter M and Opp KD (2001) Introduction. In *Social norms*, ed. by Hechter M and Opp KD, pp. xi-xx. Sage: New York.

Heckathorn DD (1989) Collective action and the second-order free-riding problem. *Rationality and Society* 1: 78-100.

Hemenway, D (2013) Preventing Gun Violence by Changing Social Norms. *JAMA Internal* Medicine 173: 1167-1168.

Hevenstone D (2015) *The American Myth of Markets in Social Policy: Exacerbating Inequality?* New York: Palgrave McMacmillan.

Hofstede G (2001) *Culture's consequences. Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks: Sage.

Horne K (2004) Collective benefits, exchange interests, and norm enforcement. *Social Forces* 82: 1037-62.

Horne K (2007) Explaining norm enforcement. *Rationality and Society* 19: 139-170.

Janssen MA and Bushman C (2008) Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology* 254: 541-45.

Keizer K, Lindenberg S, and Steg L (2008) The spreading of disorder. *Science* 322: 1681-85.

Kelling GL and Coles C (1996) *Fixing Broken Windows: Restoring Order and Reducing Crime in Our Communities*. New York: Free Press.

Kelling GL and Sousa WH Jr (2001) *Do Police Matter? An Analysis of the Impact of New York City's Police Reforms*. New York: Center for Civic Innovation at the Manhattan Institute.

Keuschnigg, M and Wolbring T (2015) Disorder, social capital, and norm violation: Three field experiments on the broken windows thesis. *Rationality and Society* 27: 96-126.

Levitt SD and List JA (2007) What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives* 21: 153-74.

Milgram S, Liberty HJ, Toledo R and Wackenhut J (1986) Response to intrusion into waiting lines. *Journal of Personality and Social Psychology* 51: 683-89.

Oliver P (1980) Rewards and punishments as selective incentives for collective action: theoretical investigations. *American Journal of Sociology* 85: 1356-75.

Orne MT (1962) On the social psychology of the psychological experiment: With particular

reference to demand characteristics and their implications. *American Psychologist* 17: 776-83.

Posner, RA and Rasmusen, EB. (1999) Creating and enforcing norms, with special reference to sanctions. *International Review of Law and Economics* 19: 369–382.

Przepiorka, W and Berger J (Forthcoming) The sanctioning dilemma: A quasi-experiment on social norm enforcement on the train.

Sampson RJ, and Raudenbush SW (1999) Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American Journal of Sociology* 105: 603-651.

Schmitt BH, Dubé L and Leclerc F (1992) Intrusions into waiting lines: does the queue constitute a social system? *Personality and Social Psychology* 63: 806-15.

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

Willer R (2009) The status solution to the collective action problem. *American Sociological Review* 74: 23-43.

Wilson JQ and Kelling GL (1982) Broken windows: The police and neighborhood safety. *The Atlantic Monthly* (March): 29–39.

Wolbring T, Bozoyan C and Langner D (2013) Walk left, stand right! A field experiment on the enforcement of informal norms on escalators. *Zeitschrift für Soziologie* 42: 239-58.

Yamagishi T (1986) The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51: 110-16.

## Appendix

**Table A1.** Multinomial logits predicting the likelihood of direct sanctioning, picking up the bottle, and subtle sanctions, as compared to no response.

| | | Any First Reaction | Any Second Reaction | Target Reactions |
|---|---|---|---|---|
| *direct sanction* | | | | |
| | Bern | 0.19 | 0.77 | -3.10** |
| | NYC | -0.96* | -16.6 | -0.86 |
| | farBin | -0.69 | -15.91 | -0.99 |
| | litterTreat | -1.36** | 0.15 | 0.15 |
| | density | 0.24 | 1.52 | 0.64* |
| | density2 | -0.01 | -0.14 | -0.04* |
| | rushhour | -0.31 | 0.3 | -0.5 |
| | out-group | | | -1.6 |
| | young adult | | | 0.8 |
| | older adult | | | 2.05** |
| | male | | | -0.93 |
| | constant | -2.41 | -7.07 | -3.99 |
| *picking up the bottle* | | | | |
| | Bern | -0.66 | 2.57* | -1.30* |
| | NYC | -1.22*** | 0.32 | -2.17** |
| | farBin | -1.79*** | -15.16 | -1.40* |
| | litterTreat | -2.53*** | 0.2 | -15.13 |
| | density | -0.08 | 0 | 0.16 |
| | density2 | 0 | 0 | -0.01 |
| | rushhour | -0.25 | -0.14 | -0.17 |
| | out-group | | | -0.43 |
| | young adult | | | 0.87 |
| | older adult | | | 0.85 |
| | male | | | -0.4 |
| | constant | -0.28 | -4.2 | -2.2 |
| *subtle sanction* | | | | |
| | Bern | 1.44*** | 1.3 | 1.56*** |
| | NYC | 0.88* | -1.06 | 1.59*** |
| | farBin | -0.63 | 0.91 | -0.58 |
| | litterTreat | -0.65** | 0.48 | -0.49 |
| | density | -0.03 | -0.05 | -0.03 |
| | density2 | -0.00 | -0.02 | 0 |
| | rushhour | -0.42 | -1.21* | -0.44 |

| | | | | |
|---|---|---|---|---|
| | out-group | | | -1.12*** |
| | young adult | | | 0.03 |
| | older adult | | | 0.29 |
| | male | | | 0.4 |
| | constant | -1.81 | 1.43 | -2.6 |
| obs | | 488 | 156 | 488 |
| LR Chi2 | | 88.47 | 49.09 | 93.18 |
| Pseudo R2 | | 0.0941 | 0.3067 | 0.146 |

**= .1 significance, ** = .05, *** = .01,*

*Caption*: multinomial logits predicting the likelihood of direct sanctioning, picking up the bottle, and subtle sanctions, as compared to no response