NORMS OF PUNISHMENT: EXPERIMENTS WITH STUDENTS AND THE GENERAL POPULATION

STEFANIA BORTOLOTTI, MARCO CASARI and FRANCESCA PANCOTTO*

Norms of cooperation and punishment differ across societies, but also within a single society. In an experiment with two subject pools sharing the same geographical and cultural origins, we show that opportunities for peer punishment increase cooperation among students but not in the general population. In previous studies, punishment magnified the differences across societies in people's ability to cooperate. Here, punishment reversed the order: with punishment, students cooperate more than the general population while they cooperate less without it. (*JEL* C72, C90, Z13)

I. INTRODUCTION

The issue of external validity of laboratory experiments has received increasing attention in the last decades. While the vast majority of experiments are conducted with a fitting sample of college students, it remains an open question whether the behavior observed in such studies is informative about society at large. Here we focus on experiments on *social dilemmas* to study other-regarding preferences and civic norms of cooperation. We compare cooperation levels in two distinct subject pools originating from the same geographical area. One sample was drawn from the student population of a large, public university (*Student* treatment). The other sample was drawn from the general adult population (*Representative* treatment) and stratified according to gender, age, and employment status. Everyone participated in public good games with and without opportunities for peer punishment.

There are many contributions in the literature that compare the behavior of students with other pools of participants with the explicit aim to test the external validity of laboratory experiments.[1] The existing evidence appears to suggest that students are less prosocial than other subject pools. For instance, in a prisoner's dilemma, students cooperate less than white-collar workers (Bigoni, Casari, and Camera 2012) or bicycle messengers (Burks, Carpenter, and Goette 2009). Similarly, students are less prosocial than rural and urban citizens in a public good game (Gächter and Herrmann 2011), than rural villagers in the appropriation of common-pool resources (Cardenas 2005), and than employees in a dictator game

*This paper is part of a larger research project that includes also Maria Bigoni and Diego Gambetta, who contributed to the design of the experiment with the general population. We thank Maria Bigoni also for her active help in running the sessions and programming the software. The authors thank Michele Belot, Peter Martinsson, participants in the seminar held at the University of Modena and Reggio Emilia, the IMEBE 2013 meeting in Madrid, the 2013 Firenze Experimental Economics workshop, and the 2014 AEW meeting in Rome for their helpful comments and suggestions on previous versions of this paper. We gratefully acknowledge the financial support of the ERC Starting Grant Strangers 241196. The usual disclaimer applies.

Bortolotti: Research Fellow, Department of Economics, University of Bologna, Bologna 40126, Italy. Phone +39 051 209 8135, Fax +39 051 209 8143, E-mail stefania.bortolotti@unibo.it

Casari: Professor, Department of Economics, University of Bologna, 40126 Bologna, Italy. Phone +39 051 209 8662, Fax +39 051 209 8493, E-mail marco.casari@unibo.it

Pancotto: Associate Professor, Department of Communication and Economics, University of Modena and Reggio Emilia, Reggio Emilia 40121, Italy. Phone +39 0522 523 264, Fax +39 0522 523 205, E-mail francesca.pancotto@unimore.it

1. The categories studied include business people and managers (Alpert 1967; Cooper 2006; Cooper, Lo, and Gu 1999; Croson and Donohue 2006; Dejong, Forsythe, and Uecker 1988; Fehr and List 2004); prisoners (Block and Gerety 1995); lay people (Glaser, Langer, and Weber 2005); children (Harbaugh, Krause, and Berry 2001; Harbaugh, Krause, Vesterlund 2002; Murnighan and Saxon 1998); finance industry professionals (Alevy, Haigh, and List 2007; Sade, Schnitzlein, and Zender 2006); and public affair officials (Potters and Van Winden 2000).

ABBREVIATIONS

GMM: Generalized Method of Moments
MPCR: Marginal Per Capita Return
OLS: Ordinary Least Squares
PGG: Public Goods Game

(Carpenter, Burks, and Verhoogen 2005; Dragone, Galeotti, and Orsini 2013). The gap remains when one compares students and *professionals*, that is self-selected subjects with a high degree of expertise who ordinarily deal with situations resembling the experimental task. The evidence includes studies on: voluntary contributions to a public good among elected officials (Butler and Kousser 2013) and shrimp fishermen (Carpenter and Seki 2011); threshold public goods without refunding among nurses (Bram Cadsby and Maynes 1998); trust games among CEO principals (Fehr and List 2004).[2]

More in general, the most appropriate pool of participants should depend both on the task and the goal of the study. For instance, contractors may be a better sample than college students for the external validity of auction experiments (Dyer, Kagel, and Levin 1989) and villagers may be a better sample than city dwellers for experiments about the management of a renewable natural resource (Cardenas 2005). Contractors and villagers are more appropriate than students in this case because they are more familiar with the experimental task and their behavior is more relevant because they are those who actually make the decisions in the field. Both aspects boost the external validity of the experimental results.[3]

The goal of the study is also relevant. When studying issues of bounded rationality, for instance, one may prefer participants with very high or very low cognitive skills, depending on the initial conjecture to be tested. High-score participants may be preferred when collecting evidence about the presence of a bound to rationality. Showing that game theorists choose numbers in a guessing game away from the Nash equilibrium prediction provides more compelling evidence about the descriptive inaccuracy of the theory than using a representative sample of the general population (Camerer 2003). Conversely, participants with low cognitive skills who succeed at a task suggest that such task is not too demanding. In short, to enhance the external validity of experimental results, one should recruit participants with a bias against the initial conjecture.

As we are interested in norms of cooperation and punishment of the society broadly defined, the most appropriate pool of participants would be a representative sample of the population at large. One reason is that civic norms of cooperation are likely to be an emergent property of a society. Let's consider, for instance, a society made up of young and old citizens. Cooperation and punishment behaviors can develop in different ways because of two driving forces: first, young and old citizens may follow different group-specific norms; second, the same individuals may behave differently when facing only people from the same age group or when interacting in a mixed group. For instance, youngsters may follow one norm when interacting with peers, but they may behave differently when they interact with elderly people. When deciding whether to punish or not, a young person may have no hesitations if the target is another young person (i.e., in-group), but she may refrain from punishing an old person (i.e., out-group). The propagation of group-specific norms in a society can depend both on the relative size of each group and on the interaction of in- and out-group norms. Hence, the civic norms of a society cannot be reduced to the sum of the behavior of specific subsamples.

Economic experiments conducted with a representative sample of the population are rare. Recruiting such samples is indeed a hard task because of logistic and technical issues. In addition, payments must be higher to compensate participants' opportunity costs. Table 1 summarizes experiments comparing students and a sample of the general population.[4]

We contribute to the current literature by ensuring high methodological standards and comparability across participant pools. Following Harrison, Lau, and Williams (2002), Bellemare and Kroger (2007), Falk, Meier, and Zehnder (2012), and Cappelen et al. (2010), we specified ex-ante stratification variables and quotas. To the best of our knowledge, this study, along with Cappelen et al. (2010), is the only experiment conducted in a laboratory to compare a student sample with a stratified sample. We conducted the experiment by following the same procedures for both students and the general population. Finally, in order to increase comparability across subsamples, our study, together with Falk, Meier, and Zehnder (2012), restricts participation of the student and representative samples to those subjects resident in a given region.

2. For a general review of experiments beyond social dilemmas that compare students with subject pools of professionals see Fréchette (2009).

3. The choice is task specific, as contractors would not be the most appropriate sample for studying the management of a renewable natural resource.

4. Papers comparing students and nonstudents with the aim of controlling for self-selection in the participation in experiments are beyond the scope of this paper (for a review, see Exadaktylos, Espin, and Branas-Garza 2013).

TABLE 1
Studies Comparing College Students with the General Population

	Stratified Sample	Same Procedures for Both Samples	Lotteries	Dictator	Ultimatum	4 PGG	PGG w/pun	Trust Game	Beauty Contest	Country
Two distinct samples										
This study	Y	Y(lab)	1			R	R			Italy
Cappelen et al. (2010)	Y	Y(lab)		1				R		Norway
Bellemare and Kroger (2007)	Y	N						1		Netherlands
Falk, Meier, and Zehnder (2012)	Y	Y(mail)						1		Switzerland
Carpenter, Connolly, and Myers (2008)	N*	N		1						United States
Gächter, Hermann, and Thöni (2003)	N*	N					1			Russia
Gächter, Hermann, and Thöni (2004)	N*	N				1				Russia
Belot, Duch, and Miller (2010)	N	Y(lab)	1	1		R		1	1	UK
Bosch-Domenech et al. (2002)	N	N							1	Germany/US/ Spain
Only one sample										
Harrison, Lau, and Williams (2002)	Y	Interview	1							Denmark
Ermisch et al. (2009)	Y	Interview						1		Britain
Exadaktylos, Espin, and Branas-Garza (2013)	N*	Interview		1	1			1		Spain
Bellemare, Kröger, and Van Soest (2008)	N*	Internet		1	1					Netherlands
Bellemare, Kroger, and Van Soest (2011)	N*	Internet			1					Netherlands
Dohmen et al. (2008)	N*	Interview						1		Germany
Egas and Riedl (2008)	N*	Internet				R	R			Netherlands

Notes: We consider a sample to be *stratified* if it has been selected according to prespecified categories and target quotas. N* indicates a representative sample that has not been selected ex ante according to target quotas. In the cells relative to each task, 1 indicates a one-shot game and R a repeated game.

Two more technical issues emerge when running experiments outside the group of college students. First, there can be logistical challenges when running multiple rounds. Unlike most of the previous studies that focused on one-shot experiments, we collected repeated measures of cooperation — with and without punishment — to investigate whether differences in contribution norms evolve over time or remain stable. One-shot experiments may capture the initial other-regarding disposition but not the reaction to others' choices. The second issue is the subjects' level of understanding of the rules of interaction. Uneducated participants may struggle to grasp a situation described in a formal and abstract manner. Instructions that suit well a college audience may be obscure to ordinary people. Thus, the misunderstanding of instructions may be then responsible for the behavioral differences across subject pools.

We report three main findings. First, without punishment, in line with previous evidence, we found that the general population cooperated more than college students. Second, these results do not survive the introduction of peer punishment, the introduction of the opportunity to punish increased cooperation among college students but not in the general population. Third, this result did not stem from lack of punishment as the general population sample punished more than the student sample. Punishment did not promote cooperation among the general population because it was frequently directed toward cooperators rather than free-riders. Previous studies have shown that there exist wide variations in punishment norms across societies: peer punishment opportunities enable some societies to overcome collective action problems, whereas lead other societies into feuds and revenge that harm cooperation (Henrich et al. 2010; Herrmann, Thöni, and Gächter 2008; Ostrom,

Walker, and Gardner 1992). Here we show experimentally that, even within the same culture, punishment has a beneficial or a detrimental effect on cooperation depending on the subsample of the population involved. The remainder of the article is organized as follows. Section II describes the characteristics of the subject pools, the experimental tasks and procedures; Section III presents the main results on cooperation and punishment; finally Section IV discusses results and concludes.

II. PARTICIPANTS AND DESIGN

The experiment comprises two treatments—*Representative* and *Student*—that vary only according to the composition of the participant pool. All participants, regardless of the treatment, were born within the Emilia-Romagna region (Italy). This information was common knowledge and could help subjects to form more accurate expectations about norms and the others' behavior. This present restriction was explicitly stated during the recruitment process and publicly announced by the experimenter at the beginning of each session.

The *Representative* sample was recruited among the general adult population by two professional companies, both unaware of the goal of the research. The companies contacted people by phone — both through telephone directories and private databases and a local recruiter. Recruiters were provided with a script to approach potential participants.[5] To be eligible, subjects had to: (a) be at least 18 years old; (b) be born within the province of Ravenna[6]; and (c) be resident within the province of Ravenna. The sample was stratified according to age (18–39, 40–59, 60 or older), sex, and employment status (employed, homemakers or retired, others — including students and unemployed). The target quotas for each category were defined according to the composition of the Italian population.[7] To favor wider participation, the subjects received a 30 Euros fuel voucher as show-up fee in addition to the earnings gained through the sessions.

The *Student* sample was recruited among the students of the University of Bologna. The University of Bologna has around 90,000 students with campuses in four of the eight provinces of the region Emilia-Romagna. Only students that were born in Emilia-Romagna were invited and could take part in the study.[8] Invitations were sent to subjects present in the ORSEE (Greiner 2004) database of the Bologna Laboratory for Social Sciences (BLESS) at the time of the experiment.[9] This sample comprises a standard participant pool of college students, which is roughly balanced between Humanities, Science, and Economic and Business majors.[10]

Table 2 reports the sociodemographic characteristics of the two samples. While gender composition is similar, age and employment compositions differ widely. In the *Student* sample the vast majority of participants is aged between 18 and 39, whereas in the *Representative* sample most participants belong to the 40 – 59 category (44.7%) and the remaining subjects are equally distributed across 18–39 and 60 or above. About half of the participants in the *Representative* sample are employed and about 13% are students. The overwhelming majority of participants self-reported in the questionnaire to be at least second-generation natives of the region. Participants share deep-rooted geographical origins, which

5. For a detailed description of the recruitment process, see the Appendix.

6. Ravenna is one of the eight provinces of Emilia-Romagna.

7. These data were collected as part of a wider research project to investigate social norms across various locations in Italy (Bigoni et al. 2013), where Ravenna was selected as one of the provinces of interest. For sample stratification, we referred to the figures of the National Institute of Statistics concerning inhabitants in January 1, 2009 (source Istat: http://demo.istat.it/pop2009/index1.html).

8. Because of the limited number of students that were born within the province of Ravenna and present in the ORSEE database, we decided to include among the potential participants subjects that were born in all the provinces of Emilia-Romagna: all of them shared similar socio-economic characteristics. As pointed out by Harrison and List (2004), there are at least two factors that may restrict the generalizability of laboratory results obtained with students: (a) there is an endogenous sample selection among students participating in experiments; (2) students are not informative about the general population. As we are mainly interested in (2), we do not take any additional precaution to limit endogenous sample selection among students. In the same spirit, we did not exclude from the database the small proportion of nonstudents that used to take part in experiments, we however retain the term "Student" for brevity.

9. As we are mainly interested in assessing to what extent results obtained with a standard participant pool can be extended to the general public, we opted for a group of participants as similar as possible to the one commonly involved in standard lab experiments. To this end, we sampled participants from the ORSEE database rather than from the general college population of the University of Bologna.

10. The database includes a small fraction of nonstudents, most of them were former students living in the area. We had 18 nonstudent participants (17%). Thanks to a questionnaire we know that 14 are 32 years old or younger and that 9 hold a college degree. About 1/3 of them hold a college degree and are looking for their first job.

TABLE 2
Sociodemographic Characteristics of the Two Samples

	Representative Sample	Student Sample
Male	51.5%	55.8%
Age		
18–39	24.3%	95.2%
40–59	44.7%	4.8%
60 or above	31.1%	0.0%
Employment status		
Employed	47.6%	8.6%
Unemployed	10.7%	7.7%
Students	13.6%	82.7%
Housewife or retired	28.2%	1.0%
Education level		
8th grade or lower	18.5%	1.0%
High school	47.5%	55.8%
College, Master, or PhD	34.0%	43.3%
Rootedness		
Elementary school in the region (county)	86.4%	97.1%
Mother born in the region (county)	69.9%	72.1%
Father born in the region (county)	63.1%	70.2%
Sessions		
Dates (dd/mm/yyyy)	02/03/2011 04/03/2011 05/03/2011 01/10/2011	23/02/2011 24/03/2011 24/03/2011 16/06/2012
No. of participants	108	104

Notes: Self-reported answers from a post-experimental computerized questionnaire. Owing to a software failure, questionnaire answers for one *Representative* session (02/03/2012) were collected via phone a few weeks after the session. Five participants did not answer the phone; as a result, for the representative sample, questionnaire data are available for 103 of 108 subjects.

may suggest shared social norms: as a matter of fact, about 87 (84)% of the participants in the *Representative* sample (*Student* sample) have one or both parents born in the region.[11]

Each session included a series of repeated Public Goods Games (PGG) with and without punishment (within-subjects design).[12] Tasks were presented in a fixed order in all sessions: each subject first played 8 periods of a PGG-Standard and then 8 periods of a PGG-Punishment. We followed this order to help the general population to better understand the punishment mechanism, which could have been more difficult to grasp had it been presented first.[13] Before each period, participants were divided into groups of $N = 4$ under a strangers-matching protocol. Interaction was anonymous and there was no possibility to build an individual reputation: a subject could not verify whether the same participant was in his/her group in the following periods.

In the PGG-Standard, each subject received an endowment of $w_i = 20$ tokens and had to decide simultaneously how to allocate those tokens between a group account (x) and a private account ($w_i - x$). Each group comprised $N = 4$ members and contributions to the group account could only take four levels, $x_i = \{0, 6, 14, 20\}$. Individual earnings were determined as follows:

$$\pi_i^1 = w - x_i + a\sum_{j=1}^{N} x_j$$

where the marginal per capita return (MPCR) of the public good was $a = 0.5$. At the end of each period, a subject could observe individual contributions and earnings for each group member. Earnings cumulated from one period to the next. The PGG-Punishment was identical to the PGG-Standard but for the addition of a second stage in which subjects had the opportunity to reduce, at a cost, the earnings of the other group members. After receiving feedbacks on individual contributions, every subject could assign $p_i = \{0, 1, 2\}$ deduction points to each group member; a deduction point had a cost of 1 token for the punisher and reduced the earnings of the targeted subject by $b = 4$ tokens. Punishment decisions were simultaneous and earnings were computed as follows:

$$\pi_i = \pi^1 - b\sum_{j \neq i}^{N} p^i - \sum_{j \neq i}^{N} p^j.$$

At the end of each period, a subject could observe the deduction points he/she received and his/her final earnings. The punisher's identity was not revealed.

In a one-shot interaction, it is a dominant strategy for rational self-interested subjects to contribute zero in both PGG-Standard and PGG-Punishment, because the marginal per capita

11. The figures for the representative sample refer to the province of Ravenna.

12. Each session included a total of five parts presented in a fixed order: (1) choice over lotteries; (2) PGG-Standard; (3) PGG-Punishment; (4) PGG-Standard; (5) PGG-Threshold. Subjects received a feedback on part 1 only at the end of the session. For the comparison of norms of cooperation across subject pools, we focus only on parts 2 and 3. Instructions for all five parts are in the Appendix.

13. We did not control for order effect. Previous studies with a similar set-up found no significant evidence of order effect (see Herrmann, Thöni, and Gächter 2008, 5, SOM).

return of the public good is below 1 and above $1/N$, and to assign zero deduction points in PGG-Punishment. Group surplus is instead maximized when everyone contributes their whole endowment and never punishes.

The study comprised eight experimental sessions, equally divided across treatments for a total of 212 subjects. Participants in a session ranged between 20 and 32 and the laboratory hardware and set-up were identical across subject pools and locations. The same experimenter read the instructions in all sessions. *Representative* sessions were held in Faenza in a large hotel conference room in the city center, where we deployed the mobile BLESS. *Student* sessions took place in Bologna at the permanent BLESS laboratory.[14]

In an effort to make the task more intuitive, we largely relied on graphical elements.[15] To facilitate elderly people unfamiliar with computers, all choices could be made by simply touching the screen (see sample screens in Appendix S1, Supporting Information) and there was indeed no need to type or use a mouse. At the end of the session, subjects filled in a questionnaire. The average *Student* (*Representative*) session lasted about 90 (120) minutes. Subjects were paid in private at the end of the session. The experiment paid 1 Euro for every 40 tokens earned. There was no show-up fee in the *Student* sessions and a 30 Euros fuel voucher in the *Representative* sessions, under the assumption of a lower opportunity cost for students than for the general adult population. Average per-capita earnings were 19.50 Euros in the *Student* sessions and 17 Euros (plus the show-up fee) in the *Representative* sessions.

III. RESULTS

We report five main results; we first consider aggregate behavior (Results 1, 2, and 3) and then present the evolution of contributions and punishment norms over time (Results 4 and 5).

In the PGG-Standard, how do observed contribution levels in the student population compare to the ones observed in the representative population?

RESULT 1. *The representative sample cooperates more in the standard Public Goods Game than the student sample.*

The average cooperation level over the eight periods was 9.1 in the *Representative* and 6.8 in the *Student* treatment. Support for Result 1 is provided by Figure 1 and an ordered logit regression, where the dependent variable is the contribution level of a subject in a period (Table 4, Model 1).[16] The main explanatory (dummy) variable *Representative sample* has a positive and highly significant coefficient, hence suggesting that the general public cooperates more than college students. To account for subjects' understanding, we also included the dummy *Low understanding* that takes into account the number of mistakes in the control questions and the time used to answer correctly to all questions. The dummy takes value 1 for subjects in the last decile of the distribution according to either the number of mistakes or the total answering time. Our results are robust to alternative ways to model understanding: in Model 2 we included a dummy that takes value 1 for subjects who made *4 or more mistakes* in the control questions and 0 otherwise. While subjects who made more mistakes contribute significantly more in the PGG-Standard, the difference between student and representative sample remains large and significant.[17]

When following a very conservative approach and considering each session as an independent observation, the difference in contributions across subject pools in PGG-Standard is not statistically significant (Mann–Whitney rank-sum, $p = .149$, $N_R = N_S = 4$, two-sided).

14. Upon arrival, subjects were seated at a visually separated desk; no form of communication was allowed during the experiment. A paper copy of the relevant instructions was handed out before each part and read loud by the experimenter. Before PGG-Standard and PGG-Punishment, subjects had to answer a computerized quiz to ensure their understanding. Everyone had to answer all questions correctly before proceeding. The experiment was programmed and conducted with the software z-Tree (Fischbacher 2007).

15. In programming our interfaces, we took inspiration from the first wave of experiments conducted at the Internet Laboratory for Experimental Economics, iLEE (for further details see: http://www.econ.ku.dk/cee/ilee/description/ilee1/).

16. We opted for ordered probit regressions to take into account that the dependent variable was not continuous but could take on only four values. Models were estimated using the Gllamm package (http://www.gllamm.org/). We also run OLS specifications and Tobit models to account for censoring at 0 and 20. Our results are robust to the use of these different estimation procedures. Results of these additional estimations are available upon request from the authors.

17. In addition, we control for three alternative ways of modeling low understanding: (1) subjects in the last quartile of the distribution according to either the number of mistakes or the total answering time; (2) subjects without a college degree or higher; (3) subjects who contributed 6 or 14 in the PGG-Threshold. Results are qualitatively similar under all specifications and are available upon request from the authors.

FIGURE 1
Contributions to the Public Good Over Time

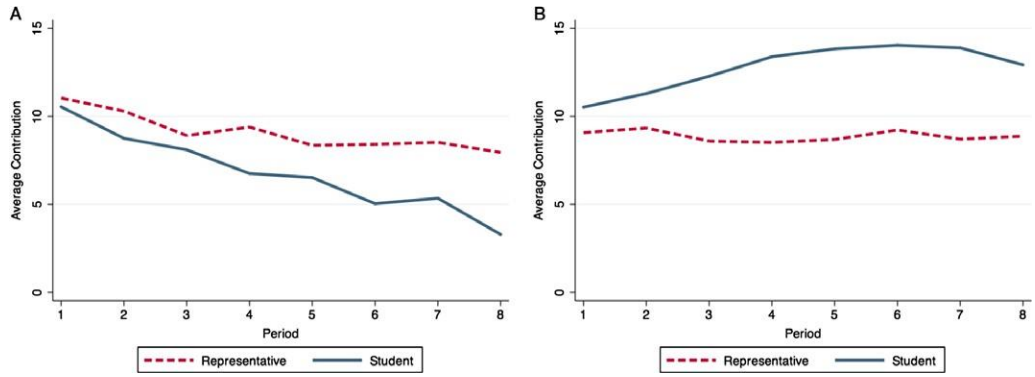
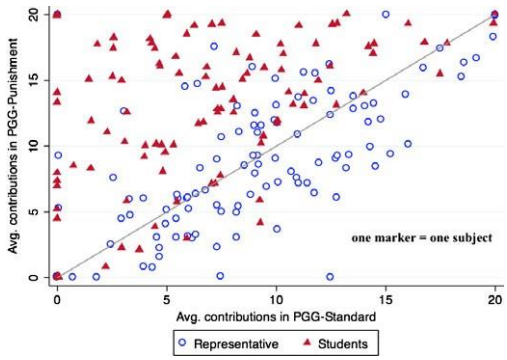


FIGURE 2
Average Individual Contributions in PGG-Standard and PGG-Punishment



In PGG-Punishment, how do contribution levels observed in a student population compare to contributions in the general population?

RESULT 2. *With punishment, the representative sample cooperates less than the student sample. The opportunity of peer punishment enhances cooperation levels in the student sample but not in the representative sample.*

The introduction of peer punishment reverses the treatment order: the general population contributes less as compared to students. Average cooperation in the PGG-Punishment was 8.9 in the *Representative* and 12.8 in the *Student* treatment. Support for Result 2 comes from Table 4 and Figure 2.

The difference across subject pools in the PGG-Punishment is highly significant according to an ordered logit regression on individual contributions (Table 4, Model 4). The negative coefficient of the explanatory variable *Representative sample* lends support to the evidence that students are more cooperative than the general public in the PGG-Punishment. The difference is also statistically significant according to a Mann – Whitney rank-sum test (p = .021, $N_R = N_S = 4$, two-sided). Moreover, the opportunity of peer punishment enhances cooperation levels in the student sample but not in the representative sample (Mann–Whitney signed-rank test, p = .068, $N_{S\text{- PGG- std}} = N_{S\text{ - PGG - punish}} = 4$, p = .465, $N_{R\text{- PGG- std}} = N_{R\text{ - PGG - punish}} = 4$, two-sided).[18] To illustrate this outcome, we plotted individual average contributions in the two variants of the PGG by subject (Figure 2). About 82% of students contribute on average more with than without punishment opportunities (vs. 32% in the representative sample). The upward shift in students' contributions is present for free-riders and contributors alike.

We are going to consider individual decisions over time in order to grasp a better understanding of the underpinning dynamics of cooperation. As a matter of fact, our experiment offers repeated measures of cooperation; this allows us to analyze the initial contribution levels as well as the

18. Per-period profit decreases from PGG-Standard to PGG-Punishment for both subject pools. In the *Representative* treatment the earnings drop was more pronounced; subjects earned, on average, about 9.1 tokens less in each period. The loss was of only 1.5 tokens among students.

TABLE 3
Average Contributions to the Public Good

	All Periods		First Period	
PGG	Representative	Student	Representative	Student
Standard	9.11	6.79	11.04	10.54
Punishment	8.88	12.77	9.07	10.52

Note: Average individual contributions to the public good, divided by subject pool and stage game.

dynamics of contribution and punishment over time (Figure 1).

RESULT 3. *Cooperation in the initial period is indistinguishable between representative and student samples both with and without opportunities to punish.*

Table 3 and Figure 1 provide support for Result 3. In the PGG-Standard, individual contributions in the first period are not significantly different across subject pools (Mann – Whitney rank-sum, $p = .614$, $N_R = 108$, $N_S = 104$, two-sided). The same conclusion holds for the PGG-Punishment ($p = .169$, $N_R = 108$, $N_S = 104$, two-sided).[19] We also regressed contributions in the first period over the dummy *Representative sample* (see Table A1 in Appendix) and it turns out that differences across treatments are not statistically significant for both PGG-Standard (Model 1) and PGG-Punishment (Model 2).

As shown in Figure 1, differences across treatments emerged over time. While in the first period, the two pools are indistinguishable, in the last period of the PGG-Standard, the representative sample shows a cooperation level more than twice as large as the student sample (7.9 and 3.3, respectively). In particular, cooperation among students unravels rather quickly, whereas the general population manages to sustain a more stable contribution level. Support for this finding is provided in Table 4 (Model 3). The negative coefficient for *Period* reasserts the presence of a declining trend in the PGG-Standard, whereas the positive coefficient in the interaction term indicates that the decline in the *Representative* treatment is less pronounced than in the *Student* treatment. The dynamics in the PGG-Punishment were exactly the opposite (see Model 6); contributions tend to increase over time and the upward trend is more marked in the *Student* than in the *Representative* treatment.

What drives these different trends in cooperation across games and subject pools? To answer this question, in the last part of this section we will focus on individual decisions to contribute and punish. We first consider whether the reaction to others' contributions—that is, conditional cooperation—is the same across treatments. Are the adjustment dynamics the same in our two participant pools?

RESULT 4. *In the representative sample, current contributions depend less on observed past contributions than in the student sample.*

We consider an indirect measure of conditional cooperation (Fischbacher, Gächter, and Fehr 2001; Kocher et al. 2008) and test how current contributions adjust to previous contributions made by others.[20] Here we mostly focus on the PGG-Standard that in our view provides a cleaner test of conditional cooperation. Indeed in the PGG-Punishment previous contributions are likely to be connected with punishment and not just with cooperative behavior.[21]

Table 5 (Models 1 to 3) lends support to Result 4 for the PGG-Standard. In all specifications, the dependent variable is the contribution level at time t for each subject. In the first two models we consider each sample separately and the regressor of interest is the sum of other group members' contributions in period $t - 1$ (*Others' contributions in* $t - 1$).[22] In PGG-Standard, *Others' contributions in* $t - 1$ has a positive and highly significant impact on the student sample but is not significant in the representative sample (Models 1 and 2, respectively in Table 5).

19. For first period data, we consider each subject an independent unit of observation.

20. Conditional cooperation is commonly defined as the willingness to contribute to the common pool based on the expectation that others will contribute as well. We consider an indirect measure and assume that a subject's belief about future group members' contributions depends on their past contributions. Our strangers-matching protocol weakens this relation compared to a partner-matching protocol. Alternatively, one could have used the strategy method to directly elicit conditional cooperation.

21. If high cooperators are more likely to punish than free riders, there should be a correlation between the punishment received by a subject and others' contributions in the previous period.

22. We also control for time trend and low understanding as in Table 4.

TABLE 4
Treatment Effect on Contributions

	Dependent Variable: Contribution					
	PGG-Standard			PGG-Punishment		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Representative sample	0.842***	0.676***	−0.117	−1.487***	−1.631***	−0.573
	(0.252)	(0.261)	(0.327)	(0.309)	(0.326)	(0.367)
Low understanding	−0.018		−0.019	−0.506		−0.527
	(0.337)		(0.362)	(0.412)		(0.421)
4 or more mistakes		0.576**			0.500	
		(0.284)			(0.352)	
Period			−0.365***			0.192***
			(0.033)			(0.032)
Period × Representative			0.237***			−0.211***
			(0.044)			(0.043)
No. of observations	1696	1696	1696	1696	1696	1696

Note: Ordered logit regression on individual contribution levels, individual-level random effects.
*** and ** indicate significance at the 1% and 5% level, respectively.

TABLE 5
Conditional Cooperation and Observed Contributions

	PGG-Standard			PGG-Punishment		
Dependent Variable: Contribution	Representative Students		Pooled Sample	Representative Students		Pooled Sample
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Others' contributions in $t-1$	0.004	0.030***	0.035***	0.021***	0.042***	0.043***
	(0.006)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)
Period	−0.099***	−0.284***	−0.173***	−0.012	0.071*	0.022
	(0.037)	(0.045)	(0.028)	(0.036)	(0.042)	(0.027)
Low understanding	0.560	−0.520	0.072	−0.284	−0.796	−0.515
	(0.475)	(0.590)	(0.370)	(0.564)	(0.710)	(0.446)
Representative sample			1.760***			−0.693
			(0.356)			(0.458)
Others' contribution in $t-1$ × Representative			−0.034***			−0.021**
			(0.009)			(0.010)
No. of observations	756	728	1484	756	728	1484

Note: Ordered logit regression on cooperation levels with individual random effects and robust standard errors (in parentheses).
***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

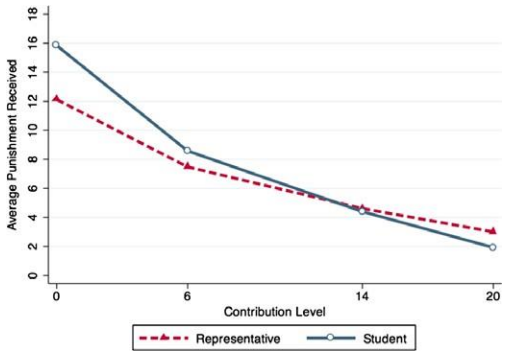
This result is confirmed also in the pooled sample (Model 3).[23]

Models 4 to 6 in Table 5 replicate the same analysis for the PGG-Punishment. Both pools tend to adjust to observed contributions. However, the difference in conditional cooperation between the two samples is less pronounced in the PGG-Punishment than in the PGG-Standard: the coefficient of interaction *Others*' *contributions in* $t-1$ × *Representative sample* is indeed negative, although not significant (see Model 6).

We now take into account the analysis of punishment behavior. The differential impact of punishment on the two subject pools may be the result of different amounts of punishment or different types of punishment. We say punishment is *prosocial* when the target of the punishment is a free-rider; conversely, we say punishment is *antisocial* when the target is a high contributor. Does the representative pool punish less than the student pool? Or does the representative pool punish differently from the student pool?

23. As a robustness check, we run the same regressions using a generalized method of moments (GMM) system methodology to check for potential endogeneity of the variable *Others*' *contributions in* $t-1$. Results are consistent with the present estimates and are reported in the Appendix.

FIGURE 3
Received Punishment by Contribution Level



RESULT 5. *The representative sample punishes no less than the student sample but engages more in anti-social punishment.*

Support for Result 5 is presented in Figure 3 and Tables 6 and 7. The extent of punishment is similar across treatments and, if anything, it is higher in the representative than in the student sample (7.2 vs. 5.9).[24] Hence, the absence of a positive effect of punishment on cooperation levels in the representative sample must stem from reasons other than lack of punishment. The data suggest an explanation based on differences in the target of the punishment as well as in the response to the received punishment.

Punishment on free-riders is heavier in the *Student* than in the *Representative* treatment (15.8 vs. 12.1 average points of punishment), whereas the opposite is true for punishment on full cooperators (1.9 vs. 3.0). These differences in punishment are statistically significant according to a logit regression (Table 6). Moreover, there is no element that points to lack of understanding as a driver of punishment (Table 6); if anything, subjects with a lower level of understanding tend to engage in more prosocial and less antisocial punishment as compared to subjects that did best in the control questions. Figure 3 illustrates this pattern. The steeper line indicates more favorable incentives for cooperation.

Another way to measure punishment preferences is the level of prosocial versus antisocial punishment. In line with other studies, prosocial punishment is more frequent than antisocial punishment but the ratio is very different in the representative and in the student sample (2.5:1 vs. 4.8:1, respectively). Notice that this treatment difference is present from the first period of interaction, which suggests that revenge is not enough to account for antisocial punishment.

Those who deviate from the average group contribution are punished significantly more, and punishment is more severe for less-than-average contributions (i.e., a negative deviation) as compared to more-than-average contributions (see Models 1 and 2 in Table 7). Sign and magnitude of these coefficients are consistent with similar studies using a strangers-matching protocol (see Fehr and Gächter 2000).[25]

24. A Wilcoxon rank-sum test does not reveal any statistically significant difference when taking each session as an independent observation ($p = .149$, $N_R = N_S = 4$).

TABLE 6
Received Punishment by Contribution Level

	Dependent Variable: Deductions Assigned (1 = Yes; 0 = No)			
	$x_i = 0$ Model 1	$x_i = 6$ Model 2	$x_i = 14$ Model 3	$x_i = 20$ Model 4
Representative sample	−1.234*** (0.454)	−0.364 (0.354)	−0.073 (0.393)	1.324** (0.517)
Low understanding	−0.837 (0.586)	−0.322 (0.473)	0.263 (0.522)	1.825*** (0.648)
No. of observations	1074	1239	1344	1431

	Dependent Variable: Deductions Assigned (1 = Yes; 0 = No)			
	$x_i = 0$ Model 1	$x_i = 6$ Model 2	$x_i = 14$ Model 3	$x_i = 20$ Model 4
Representative sample	−1.089** (0.473)	−0.404 (0.371)	−0.234 (0.416)	0.991* (0.545)
Four or more mistakes	−0.541 (0.494)	0.160 (0.397)	0.541 (0.452)	1.179** (0.583)
No. of observations	1074	1239	1344	1431

Note: Logit regression on assigned punishment, with individual-level random effects.
***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

25. Models 1 and 2 report results for *Representative* and *Student* treatments, respectively. The variable *Negative deviations* (*abs*) has a positive coefficient and is highly significant in both treatments (see Models 1 and 2) hence giving support to the idea that the more the contribution falls short of others' contributions the more severe the punishment. Quite surprisingly, also the coefficient of the variable *Positive deviations* is positive and significant. That implies that punishment increases as the gap between others' contributions and socially minded subjects' contributions widens. The negative and highly significant coefficient for *Others' contributions* implies that a deviation from others' contributions is punished more severely if the sum of the contributions is small.

TABLE 7
Treatment Effect on Punishment

Dependent Variable: Deduction Points Received	Representative Model 1	Students Model 2	Pooled sample Model 3
Other's contributions	−0.019***	−0.033***	−0.032***
	(0.006)	(0.007)	(0.007)
Positive deviation	0.104***	0.059***	0.062***
	(0.016)	(0.023)	(0.022)
Negative deviation(abs)	0.209***	0.356***	0.341***
	(0.019)	(0.022)	(0.018)
Period	−0.033	−0.019	−0.027
	(0.027)	(0.031)	(0.020)
Representative Sample			0.469
			(0.365)
Others' contrib. × Representative			0.012
			(0.009)
Pos. deviation × Representative			0.041
			(0.027)
Neg. deviation × Representative			−0.125***
			(0.024)
No. of observations	864	832	1696

Notes: Ordered logit regression on deduction points received, individual-level random effects. *Negative deviation* is the absolute value of the deviation of a subject's contribution level with respect to the average contribution of the others in her group, in the case that the contribution falls short of the average, and 0 otherwise. *Positive deviation* takes values other than 0 when a subject's contribution is larger than the average contribution of the others.

*** indicates significance at the 1% level.

When pooling all samples, the evidence suggests again that the representative sample sanctions relatively fewer free-riders and relatively more contributors than the student sample (Models 3 and 4 in Table 7). This pattern could have discouraged cooperation and might explain the weak impact of punishment within the representative sample. In the representative sample there is significantly less punishment of defectors than in the student sample (*Negative deviation* (*abs*) × *Representative*).

Besides shifting the target of punishment, the representative sample also responds weakly to punishment received. The evidence comes from logit regressions on the variations over time in contributions levels of free-riders and full cooperators (Table 8). More specifically, the dependent variable takes value 1 if the contribution level in period t is different from $t - 1$, 0 otherwise. Free-riders who receive punishment do not subsequently increase their cooperation level; and full cooperators who receive punishment do not decrease their cooperation level (Models 1 and 4). These results stand in sharp contrast to the behavior of the student sample, which strongly reacts to punishment (Models 2 and 5). The treatment differences are significant (see *Deduction received in* $t - 1$ × *Representative* in Models 3 and 6).

A comparison between the behavior of the student sample versus the young subjects in the representative sample could be of interest. If the behavior in the two groups is similar then the added value of a representative sample would mostly originate from the variety in sociodemographic characteristics. If the behavior differs then it becomes empirically relevant also how the same subject adapts his behavior depending on who the others are. A first exploratory analysis points toward the former interpretation. Given the limited number of young people within the representative sample, further studies are in order before making firm claims.

IV. DISCUSSION AND CONCLUSIONS

This study compares the cooperative behavior of two samples sharing similar geographical and cultural origins but differing along important sociodemographic dimensions: college students and a representative subsample of the general adult population. We find that results from experiments on norms of cooperation and punishment among students cannot be readily generalized to society at large.

In a social dilemma, we replicate the common finding that students in a simple collective action task are on average less cooperative than the general population (Result 1, see for instance, Bellemare and Kroger 2007; Bellemare, Kröger, and Van Soest 2008; Cappelen et al. 2010; Belot, Duch, and Miller 2010). Previous studies show that, when facing social dilemmas, some societies benefit from the availability of opportunities for peer punishment while others do not, and punishment opportunities magnify the existing differences across societies in their ability to cooperate (Herrmann, Thöni, and Gächter 2008). Here we show that, even within the same society, the impact of peer punishment in promoting cooperation can vary widely depending on the subsample of the population considered. Our results document that punishment can reverse the ordering of subgroups in a society in terms of cooperativeness even when both participant pools are from the same geographical area. In a public goods game, punishment opportunities had a positive effect on cooperation in the student subsample, whereas little or no effect was detected in

TABLE 8

Variation in Contribution Levels and Punishment: High versus Low Contributors.

Dependent Variable: Delta Contributions 1 if $\lvert Give_{(t)} - Give_{(t-1)} \rvert > 0$	Contributes 0 in $t-1$			Contributes 20 in $t-1$		
	Representative Model 1	Students Model 2	All Samples Model 3	Representative Model 4	Students Model 5	All Samples Model 6
Deduction received in $t-1$	−0.101	0.453**	0.428**	0.293	0.806***	0.953***
	(0.141)	(0.230)	(0.213)	(0.311)	(0.204)	(0.218)
Period	0.048	0.038	0.039	0.2	−0.009	0.064
	(0.103)	(0.165	(0.087)	(0.143)	(0.092)	(0.078)
Low understanding	0.491	0.085	0.324	−0.261	1.060*	0.658
	(0.747)	(1.072)	(0.612)	(1.325)	(0.623)	(0.622)
Representative sample			2.302**			2.65***
			(1.112)			(0.559)
Deduction received in $t-1$ × Representative			−0.534**			−0.747**
			(0.258)			(0.333)
No. of observations	203	111	314	136	276	412

Notes: Logit regression on variation in cooperation levels with individual random effects and clusters at the session level. The dependent variable takes value 1 if contributions in t and $t-1$ are not identical and 0 otherwise. Models 1 to 3 consider subjects who contributed 0 in $t-1$. Models 4 to 6 consider subjects who contributed 20 in $t-1$.

***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

the general population. As a consequence, without peer punishment, students contributed less than the general population; with peer punishment students were more cooperative than the general population (Result 2).

We found two main factors driving this differential effect of peer punishment. One factor lies in distinct preferences for punishment. There were differences in the way punishment was used by the two participant pools: for instance, punishment levels were higher in the *Representative* than *Student* treatment. More importantly, in the general population a remarkable amount of punishment was directed toward cooperators (i.e., antisocial punishment) and this happened with a higher frequency than in the student pool, starting from period one. Hence, punishment did not promote cooperation among the general population because it was frequently directed toward cooperators rather than free-riders. Another factor is the unresponsiveness to punishment by the general population subsample. While students, both high and low contributors, showed significant reactions to the punishment received in the previous period, those reactions were not significant in the general population. As a consequence, contributions in the student subsample increase with repetition while they remain flat in the general population.

More generally, a main behavioral difference between the subsamples is the low reactivity of the general population to the feedback within the experiment. We report no difference between the students and the general population subsamples in their first period average contribution to the public good game, either with punishment or without punishment. The differences emerge with repeated interactions. In particular, in the baseline public good game we document less conditionally cooperative behavior among the general population than among students (Result 4). In the public good game with punishment, as already mentioned, we observe a smaller reaction to past punishment within the general population than within students. One implication of this evidence is to exert caution when generalizing results of experiments consisting of one-shot social dilemmas because some differences emerge only over time.

There could be a variety of reasons for the low reactivity of the general population to experimental feedback. One reason could be the poor understanding of the rules of the experimental set-up. When venturing beyond college students, participants may lack a clear comprehension of the situation at hand. In this study, we put extra effort in the experimental design, software and instructions to facilitate understanding. Moreover, our econometric analysis supports our main results also after checking for understanding. Another possible explanation is that some participants may update their beliefs more slowly. Two motivations come to mind. A rational motivation could be past exposure to many similar experiences. A behavioral motivation is related to receiving feedback from someone

inside or outside one's own reference group. For instance, an elderly person may give low weight to the feedback of a young person, because it is deemed irrelevant.

Both motivations would suggest a slower updating in an experiment among the general population than among a homogeneous student population. There can be other reasons, such as higher cognitive costs of adjustment. In conclusion, these results should not be taken as a sweeping indictment against laboratory experiments with student populations. On the contrary, they are part of an ongoing effort to identify those research questions that can be usefully addressed using students and those that instead are best dealt with other types of participants. While students are well suited for studying a number of issues (i.e., theory testing, learning, rationality, etc.), the use of a representative sample of the general population is, in our view, the most appropriate choice when investigating the emergence and the maintenance of civic norms of cooperation and punishment, which is often the result of the interaction between different social strata. For instance, if we were to classify the Italian society according to the impact of peer punishment in promoting cooperation, one would draw opposite conclusions depending on whether the experiment was run with college students or with the general population.

APPENDIX

A. FIRST PERIOD AND DYNAMIC PANEL ESTIMATION

Result 3 suggests that contributions in the first period are indistinguishable across subject pools. In Table A1, we regress individual contributions over the dummy *Representative sample* and provide support to Result 3.

Result 4 suggests that differences emerge over time, as students condition their behavior on previous experience more than the general population does. Apart from conditional cooperation there can be two additional factors that influence cooperation: (a) individual (unconditional) preferences for cooperation; and (2) other unobserved individual characteristics. All these motivations are captured by the following equation:

$$x_{i,t} = v_i + \alpha x_{i,t-1} + \beta \sum_{j=1}^{n} x_{j,t-1} + u_{i,t} \qquad \text{(A1)}$$

where $x_{i,t}$ indicates the contribution to the public good of subject i at time t; $\sum_j x_{j,t-1}$ indicates the sum of the contributions of the other group members in the previous period and is meant to capture conditional cooperation. Please recall that groups were formed according to a strangers-matching protocol at the beginning of each period. The variable $x_{i,t-1}$ is the contribution of subject i in period $t-1$ and measures the persistence of subjects' choice; we interpret this variable as a proxy for individual preferences toward cooperation. Finally, v_i is an individual time-invariant component capturing intrinsic characteristics of each subject that cannot be observed.

TABLE A1
Table A1: Treatment Effect on Contributions in Period 1: PGG-Standard and Punishment

	Dependent Variable: Contribution in Period 1	
	PGG-Standard Model 1	PGG-Punishment Model 2
Representative sample	0.118	−0.343
	(0.298)	(0.345)
Low understanding	−0.338	−0.291
	(0.389)	(0.360)
No. of observations	212	212
Log likelihood	−284.328	−288.298

Note: Ordered logit regression on individual contribution levels in period 1, standard errors robust for clustering at the session level (in parentheses).

To account for the endogeneity problem arising from the introduction in the model of the variable $x_{i,t-1}$, we implement a two-step GMM system estimator.

Following the general model of Equation (A1) we estimate a dynamic panel of the form:

$$x_{i,t} = \alpha_1 x_{i,t-1} + \beta_1 \sum_{j} x_{j,t-1} + u_{i,t} \qquad \text{(A2)}$$

$$u_{i,t} = v_i + e_{i,t} \qquad \text{(A3)}$$

where v_i are unobserved individual effects, $e_{i,t}$ are the observation specific errors, which have zero mean ($E[e_{i,t}] = E[v_i e_{i,t}] = 0$), constant variance and are uncorrelated across time and individuals $E[e_{i,t}] \times E[e_{i,s}] = 0$ for each i, j, t, s and $i \neq j$. $\sum_j x_{j,t-1}$ is endogenous thanks to the strangers matching protocol implemented in our setting and to the fact that the choice of individual i is excluded from the calculation of the aggregate group contribution in the previous period.

Endogeneity is an issue in this context because of the small number of time periods available for the estimation, or what is defined in the literature as small sample bias (Nickell 1981). This could also be the reason for the scarce use of this methodology in the experimental literature.[26] We implement two-step GMM system estimators,[27] which are robust under heteroskedasticity with the Windmeijer (2005) finite-sample correction to avoid downward bias

26. Notable exception is Branas-Garza, Bucheli, and Garcia-Munoz (2011) that compares static and dynamic panel estimation in an experimental setting.

27. The problem of endogeneity in small samples has been originally tackled by Arellano and Bond (1991) seminal paper but other contributions have extended the applicability of the methodology in various directions in the following years. See Arellano and Bover (1995); Blundell and Bond (1998).

TABLE A2
Instruments for System GMM Estimation of PGG-Standard

Instruments for first differences equation		
Type of instrument	Standard	GMM-type
Variable used	Others' contribution	Own contribution
Specification	First difference, lag 1	Level, lags from 3 to max
Instruments for levels equation		
Type of instrument	Standard	GMM-type
Variable used	Others' contribution	Own contribution
Specification	Level, lag 1	First difference, lag 2

TABLE A3
Instruments for System GMM Estimation of PGG with Punishment

Instruments for first differences equation		
Type of instrument	Standard	GMM-type
Variable used	Others' contribution	Own contribution
Specification	First difference, lag 2	Level, lags from 2 to max
Instruments for levels equation		
Type of instrument	Standard	GMM-type
Variable used	Others' contribution	Own contribution
Specification	Level, lag 2	First difference, lag 1

As we are interested to introduce in our model a time invariant regressor, the *Representative Sample* dummy, difference GMM estimators are not appropriate (as the time invariant regressor would be canceled out in the procedure). Hence we use a system GMM estimator that maintains both the original and the differenced equation and uses both levels and differenced variables as instruments. For simplicity, the discrete dependent variable is approximated to continuous as in Hislop (1994).[28]

For each estimated Model we control the *p* values of the Sargan and Hansen test of overidentifying restrictions and the Arellano and Bond (1991) second order autocorrelation test in first differences. The former test indicates a correctly specified set of instruments, whereas the Arellano Bond test evaluates the presence of residual order autocorrelation of the differenced error, which in this context is a signal of endogeneity between the lagged endogenous variable and the differenced fixed effect, condemning the related variable to be an invalid instrument. In Tables A2 and A3, we list the instruments implemented for GMM system estimation for the Standard and Punishment treatments, respectively. In Models 5 and 6 of the Punishment treatment, a correct specification was not achieved using the same instruments: this suggests that the two subject pools are substantially different and that it is necessary to explore other variables in order to find an explanation to individual behavior in the punishment treatment.

28. We leave for future research the possibility to implement a dynamic discrete choice panel with endogenous regressors as in Stewart (2006), for example.

Table A4 reports estimates for Standard (Models 1 to 3) and Punishment (Models 4 to 6) variants of the Public Goods Game. Model 1, considers data only from the *Representative* treatment in the PGG-Standard; the large and significant coefficient of the variable *Own contribution in t* − 1 suggests that individual contributions in the general population are very persistent. On the contrary, the coefficient of the variable *Other*'*s contribution in t* − 1 is very small and not significant at any conventional level. Taken together, these two variables confirm that in the representative sample subjects tend to stick to their choices and are less influenced by others' behavior. When considering the *Student* treatment only (Model 2), we find that coefficient on *Other*'*s contribution in t* − 1 is larger and highly significant, and this supports the idea that conditional cooperation plays a key role among students even after controlling for their own contributions in *t* − 1.

When moving from the Standard to the Punishment variant, we find that none of the explanatory variables can account for observed cooperation in the *Representative* treatment (Model 4). In this case, behavior is thus explained by a variable not included in this present model; received punishment appears to be a likely candidate (see discussion in the Section III). For the student sample, the only marginally significant variable is *Other*'*s contribution in t* − 1, hence yielding further evidence in favor of the idea that students are more conditional cooperators than the general public.

B. RECRUITMENT

Recruitment Procedure for the Representative Sample

Participants to the *Representative* treatment were recruited from the general population of the province of Ravenna, which is part of Emilia-Romagna region, located in the North of Italy. Eligible candidates for the study had to: (a) be at least 18 years; (b) be born in the county; (c) be resident in the county; (d) have a good knowledge of spoken and written Italian. The experimenters, before the experimental sessions were carried out, double checked participants' ID cards so to guarantee that all subjects met the requirements (age and place of birth). At the beginning of each session, the experimenter made public that all subjects in the room were born and resident in the same province (or at least in the region) with the explicit aim to make this information common knowledge.

We wanted a representative sample of the Italian population with respect to age, sex, and employment status, as these characteristics might be relevant for the investigation of cooperation norms in a society. The sample was stratified according to three categories of age (18–39; 40–59; 60 and older), two of sex (male and female), and three for employment status (employed; housewives and retired; others, including students and unemployed). For the composition of the target sample we referred to the 2009 statistics for the Italian population.[29]

We hired two professional companies — Metis-Ricerche and Demoskopea — to recruit subjects that comply with the aforementioned requirements. We provided these companies with a message and a script to approach potential participants. Details about the study and the goal of the experiment were

29. We referred to the number of inhabitants registered on January 1, 2009. Age range: 18–39 years, 34.8%; 40–59 years, 34.6%; 60 and more, 31.6%. Sex: male, 48%; female, 52%. Employment status: employed, 42%; housewives and retired, 37%; others, 21%. Source: http://demo.istat.it/pop2009/index1.html

TABLE A4
Dynamic Panel Estimation

Dependent Variable: Contributions	PGG-Standard			PGG-Punishment		
	Representative Model 1	Students Model 2	All Sample Model 3	Representative Model 4	Students Model 5	All Sample Model 6
Own contribution in $t-1$	0.711**	0.533***	0.460***	0.059	0.143	0.109**
	(0.28)	(0.14)	(0.16)	(0.07)	(0.10)	(0.05)
Others' contribution in $t-1$	0.023	0.086***	0.051***	0.006	0.113*	0.051
	(0.03)	(0.02)	(0.02)	(0.08)	(0.06)	(0.06)
Representative sample			3.134			−6.008***
			(3.11)			(1.92)
No. of observations	756	728	1484	648	624	1272

Notes: Blundell and Bond (1998) panel estimation. *p* Value for Sargan and Hansen test (null hypothesis that the overidentifying restrictions are valid) and Arellano Bond test for second- and third-order autocorrelation in first differences: (0.686, 0.518, 0.077) for Model 1;(0.000, 0.194, 0.022) for Model 2; (0.000, 0.153, 0.016) for Model 3; (0.046, 0.178, 0.083) for Model 4; (0.002, 0.047, 0.403) for Model 5; (0.260, 0.845, 0.373) for Model 6.

***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

not disclosed to subjects during recruitment; recruiters had no prior knowledge of the purpose or the content of this study. We asked them to recruit people resident both in the town and outside the town where the experiment has been carried out; in both cases only subjects resident in the province of Ravenna could be involved. In addition to the aforementioned requirements, special categories of people were ex-ante barred from participation, such as: employees of the research sector; people that participated to market researches in the preceding 3 months; recruiters' family members; employees of marketing companies and of the press sector in general. Moreover, no more than two people per session needed to be acquainted with each other.

One company (Metis-Ricerche) recruited subjects for the first three sessions. Potential subjects were identified with the use of telephone book entries and approached by telephone calls. All phases of the recruitment process were performed from the company's headquarters, and, in case of acceptance, the company provided the participant with a confirmation letter. This letter contained the same information that was provided on the phone by the company operators and that we had previously agreed upon with the company. The former company Metis-Ricerche decided not to renew their contract for recruitment of people in other locations, as the recruitment procedures turned out to be more expensive than what they expected. The latter company (Demoskopea) was in charge of the recruitment of subjects for the last session. Local representatives of Demoskopea contacted directly subjects in each province. The local recruiters proceeded with the choice/random extraction of names from telephone books and with random contacts obtained through personal interactions as instructed by the headquarters. In the following part of our study, we report the message used by both companies to recruit subjects for our study.

Message for recruiters with instructions

We would like to invite you to participate to a meeting organized by the Universities of Bologna and Oxford. We are looking for people born in the city and within the province of Faenza. The aim of this study is strictly scientific. There are no commercial purposes and the identity of all participants will be always kept anonymous. Our interest is to understand how Italians take decisions in situations dealing with money. During the meeting you will be given several different situations and you will be kindly asked to take decisions. Taking decisions is an easy task. No particular skills are required.

We offer you a payment of 30 euros in petrol tokens, plus a sum in cash that, according to your choices and those of other participants, will amount up to 25 euros. You will be paid at the end of the meeting, which we expect to last no more than 2 hours and a half.

If you wish to verify the accuracy of these information, please contact *Name of the secretary in charge*, from University of Bologna, or visit the website http://www.unibo.it/Portale/Ricerca. If you accept our offer, you may show up *location* at *time*, which is *description of how to get to the location*.

By participating to this meeting, you make a contribution to one of the few scientific research projects supported by the European Commission in Italy.

F.A.Q. (In case somebody asks, recruiters are allowed to provide the following extra information)

- How do we make our choices?
  - Choices are made very easily, touching the screen with a finger. It is just like an ATM or a cell phone with touch screen.
- In a nutshell, how does this activity work?
  - You will be given several different situations and you will be kindly asked to choose among alternatives. There is no right answer, we just want to know your opinion.

Recruitment Procedure for the Student Sample

Subjects that belonged to the student sample were recruited according to the standard procedure implemented in a regular laboratory experiment. Announcements were sent to potential participants in the ORSEE database: this database is the one commonly used in the Bologna Laboratory for Experiments in Social Sciences (BLESS). We slightly changed the standard announcement to include the requirement that subjects to this study should be born in Emilia-Romagna. Subjects were asked to reach the laboratory

on the agreed day for the carrying out of the session with a valid ID card, which was checked by experimenters to verify the birth and residence requirements. In the following part of our study, we will provide the announcements sent via the ORSEE platform to recruit the participants for the students sessions.

Message for the student sessions

— – Please, ignore this message if you were not born in Emilia Romagna

Hello *(first name last name)*

You are kindly invited to participate to a research in our Laboratory of Experimental Economics.

Only people that were born in Emilia Romagna can take part to this study.

Please, do not sign in if you were born in another region. People born outside of Emilia Romagna, even if they sign in, will not be granted the chance to participate to the study.

Sessions wil ltake place in the following dates and time slots:

(session list)

To choose the session, please click on the following link:

(link)

(If you cannot click on the link, please select it, copy it by clicking on the right button of the mouse and paste it in the address line by clicking on right button once again.)

REFERENCES

Alevy, J. E., M. S. Haigh, and J. A. List. "Information Cascades: Evidence from a Field Experiment with Financial Market Professionals." *Journal of Finance*, 62(1), 2007, 151–80.

Alpert, B. "Non-Businessmen as Surrogates for Businessmen in Behavioral Experiments." *Journal of Business*, 40(2), 1967, 203–7.

Arellano, M., and S. R. Bond. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies*, 58(2), 1991, 277–97.

Arellano, M., and O. Bover. "Another Look at the Instrumental Variable Estimation of Error-Components Models." *Journal of Econometrics*, 68(1), 1995, 29–51.

Bellemare, C., and S. Kröger. "On Representative Social Capital." *European Economic Review*, 51(1), 2007, 183 – 202.

Bellemare, C., S. Kröger, and A. Van Soest. "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities." *Econometrica*, 76(4), 2008, 815–39.

———. "Preferences, Intentions, and Expectation Violations: A Large-Scale Experiment with a Representative Subject Pool." *Journal of Economic Behavior & Organization* 78(3), 2011, 349–65.

Belot, M., R. Duch, and L. Miller. "Who Should Be Called to the Lab? A Comprehensive Comparison of Students and Non-Students in Classic Experimental Games." Discussion Papers 2010001, University of Oxford, Nuffield College, 2010.

Bigoni, M., M. Casari, and G. Camera. "Strategies of Cooperation and Punishment among Students and Clerical Workers." *Journal of Economic Behavior and Organization*, 94, 2012, 172 – 82.

Bigoni, M., S. Bortolotti, M. Casari, D. Gambetta, and F. Pancotto. "Cooperation Hidden Fronteirs: The Behavioral Foundations of the Italian North–South Divide." Technical Report, Department of Economics, University of Bologna WP 882, 2013.

Block, M. K., and V. E. Gerety. "Some Experimental Evidence on Differences between Student and Prisoner Reactions to Monetary Penalties and Risk." *Journal of Legal Studies*, 24(1), 1995, 123 – 38.

Blundell, R., and S. Bond. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics*, 87(1), 1998, 115 – 43.

Bosch-Domenech, A., J. Montalvo, R. Nagel, and A. Satorra. "One, Two, (Three), Infinity, ...: Newspaper and Lab Beauty-Contest Experiments." *American Economic Review*, 92(5), 2002, 1687–701.

Bram Cadsby, C., and E. Maynes. "Choosing between a Socially Efficient and Free-Riding Equilibrium: Nurses versus Economics and Business Students." *Journal of Economic Behavior & Organization*, 37(2), 1998, 183–92.

Branas-Garza, P., M. Bucheli, and T. Garcia-Munoz. "Dynamic Panel Data: A Useful Technique in Experiments." Technical Report 10/22, Department of Economic Theory and Economic History of the University of Granada, 2011.

Burks, S., J. Carpenter, and L. Goette. "Performance Pay and Worker Cooperation: Evidence from an Artefactual Field Experiment." *Journal of Economic Behavior & Organization*, 70(3), 2009, 458–69.

Butler, D. M., and T. Kousser. "How Do Public Goods Providers Play Public Goods Games?" 2013.

Camerer, C. F. *Behavioral Game Theory*. New York: Russell Sage Foundation, 2003.

Cappelen, A. W., K. Nygaard, E. O. Sørensen, and B. Tungodden. "Efficiency, Equality and Reciprocity in Social Preferences: A Comparison of Students and a Representative Population." Discussion Paper Series in Economics 28/2010, Department of Economics, Norwegian School of Economics, 2010.

Cardenas, J. C. "Groups, Commons and Regulations: Experiments with Villagers and Students in Colombia," in *Psychology, Rationality and Economic Behaviour*, edited by B. Agarwal and A. Vercelli. London: Palgrave, 2005, 242.

Carpenter, J., and E. Seki. "Do Social Preferences Increase Productivity? Field Experimental Evidence from Fishermen in Toyama Bay." *Economic Inquiry*, 49(2), 2011, 612–30.

Carpenter, J. P., S. Burks, and E. Verhoogen. *Comparing Students to Workers: The Effects of Social Framing on Behavior in Distribution Games*. Bradford, UK: Emerald Group Publishing Limited, 2005, 261 – 89.

Carpenter, J., C. Connolly, and C. Myers. "Altruistic Behavior in a Representative Dictator Experiment." *Experimental Economics*, 11, 2008, 282–98.

Cooper, D. J. "Are Experienced Managers Experts at Overcoming Coordination Failure?" *The B.E. Journal of Economic Analysis & Policy*, 6(2), 2006, 1450.

Cooper, D. J., J. K. W. Lo, and Q. L. Gu. "Gaming against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers." *American Economic Review*, 89(4), 1999, 781 – 804.

Croson, R., and K. Donohue. "Behavioral Causes of the Bullwhip Effect and the Observed Value of Inventory Information." *Management Science*, 52(3), 2006, 323–36.

Dejong, D. V., R. Forsythe, and W. C. Uecker. "A Note on the Use of Businessmen as Subjects in Sealed Offer Markets." *Journal of Economic Behavior & Organization*, 9(1), 1988, 87–100.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde. "Representative Trust and Reciprocity: Prevalence and Determinants." *Economic Inquiry*, 46(1), 2008, 84–90.

Dragone, D., F. Galeotti, and R. Orsini. "Temporary Workers Are Not Free-Riders: An Experimental Investigation." Technical Report, University of Bologna, DSE-WP 915, 2013.

Dyer, D., J. H. Kagel, and D. Levin. "A Comparison of Naive and Experienced Bidders in Common Value Offer Auctions: A Laboratory Analysis." *The Economic Journal*, 99(394), 1989, 108–15.

Egas, M., and A. Riedl. "The Economics of Altruistic Punishment and the Maintenance of Cooperation." *Proceedings of the Royal Society B*, 275, 2008, 871–78.

Ermisch, J., D. Gambetta, H. Laurie, T. Siedler, and S. C. Noah Uhrig. "Measuring People's Trust." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4), 2009, 749–69.

Exadaktylos, F., A. Espin, and P. Branas-Garza. "Experimental Subjects Are Not Different." *Nature Scientific Reports*, 3(1213), 2013, 1 – 6.

Falk, A., S. Meier, and C. Zehnder. "Do Lab Experiments Misrepresent Social Preferences?" *Journal of the European Economic Association*, 11, 2012, 839–52.

Fehr, E., and S. Gächter. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, 90(4), 2000, 980–94.

Fehr, E., and J. A. List. "The Hidden Costs and Returns of Incentives—Trust and Trustworthiness among CEOs." *Journal of the European Economic Association*, 2(5), 2004, 743–71.

Fischbacher, U. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics*, 10(2), 2007, 171 – 78.

Fischbacher, U., S. Gächter, and E. Fehr. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters*, 71(3), 2001, 397 – 404.

Fréchette, G. R. "Laboratory Experiments: Professionals versus Students, " in *The Methods of Modern Experimental Economics*, edited by G. Fréchette and A. Schotter. Oxford: Oxford University Press, 2009.

Gächter, S., and B. Herrmann. "The Limits of Self-Governance When Cooperators Get Punished: Experimental Evidence from Urban and Rural Russia." *European Economic Review*, 55(2), 2011, 193 – 210.

Gächter, S., B. Hermann, and C. Thöni. "Norms of Cooperation among Urban and Rural Dwellers. Experimental Evidence from Russia." Mimeo, University of St. Gallen, Switzerland, 2003.

———. "Trust, Voluntary Cooperation, and Socio-Economic Background: Survey and Experimental Evidence." *Journal of Economic Behavior and Organization* 55, 2004, 505–31.

Glaser, M., T. Langer, and M. Weber. *Overconfidence of Professionals and Lay Men: Individual Differences within and between Tasks?* Mannheim, Germany: University of Mannheim, 2005.

Greiner, B. "The Online Recruitment System ORSEE 2.0 — A Guide for the Organization of Experiments in Economics." Working Paper Series in Economics 10, University of Cologne, Department of Economics, 2004.

Harbaugh, W. T., K. Krause, and T. R. Berry. "Garp for Kids: On the Development of Rational Choice Behavior." *American Economic Review*, 91(5), 2001, 1539–45.

Harbaugh, W. T., K. Krause, and L. Vesterlund. "Risk Attitudes of Children and Adults: Choices over Small and Large Probability Gains and Losses." *Experimental Economics*, 5(1), 2002, 53 – 84.

Harrison, G. W., and J. A. List. "Field Experiments." *Journal of Economic Literature*, 42(4), 2004, 1009–55.

Harrison, G. W., M. I. Lau, and M. B. Williams. "Estimating Individual Discount Rates in Denmark: A Field Experiment." *American Economic Review*, 92(5), 2002, 1606 – 17.

Henrich, J., J. Ensminger, R. McElreath, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, and J. Ziker. "Markets, Religion, Community Size, and the Evolution of Fairness and Punishment." *Science*, 327(5972), 2010, 1480–4.

Herrmann, B., C. Thöni, and S. Gächter. "Antisocial Punishment across Societies." *Science*, 319(5868), 2008, 1362 – 7.

Hislop, D. R. "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Woman." *Econometrica*, 6(67), 1994, 1255 – 94.

Kocher, M. G., T. Cherry, S. Kroll, R. J. Netzer, and M. Sutter. "Conditional Cooperation on Three Continents." *Economics Letters*, 101(3), 2008, 175–8.

Murnighan, J. K., and M. S. Saxon. "Ultimatum Bargaining by Children and Adults." *Journal of Economic Psychology*, 19(4), 1998, 415–45.

Nickell, S. "Biases in Dynamic Models with Fixed Effects." *Econometrica*, 49, 1981, 1417 – 26.

Ostrom, E., J. Walker, and R. Gardner. "Covenants with and without a Sword: Self-Governance Is Possible." *American Political Science Review*, 86, 1992, 404–17.

Potters, J., and F. Van Winden. "Professionals and Students in a Lobbying Experiment: Professional Rules of Conduct and Subject Surrogacy." *Journal of Economic Behavior & Organization*, 43(4), 2000, 499–522.

Sade, O., C. Schnitzlein, and J. F. Zender. "Competition and Cooperation in Divisible Good Auctions: An Experimental Examination." *Review of Financial Studies*, 19(1), 2006, 195 – 235.

Stewart, M. "Maximum Simulated Likelihood Estimation of Random-Effects Dynamic Probit Models with Autocorrelated Errors." *Stata Journal*, 6(2), 2006, 256–72.

Windmeijer, F. "A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators." *Journal of Econometrics*, 126(1), 2005, 25 – 51.