

This is the peer reviewed version of the following article:

Norms of Punishment: Experiments with students and the General Population / Bortolotti, Stefania; Casari, Marco; Pancotto, Francesca. - In: ECONOMIC INQUIRY. - ISSN 0095-2583. - STAMPA. - 53:2(2015), pp. 1207-1227. [10.1111/ecin.12187]

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

03/05/2026 07:47

(Article begins on next page)

Norms of Punishment: Experiments with Students and the General Population

Stefania Bortolotti*

University of Bologna

Marco Casari†

University of Bologna

Francesca Pancotto‡

University of Modena and Reggio Emilia

September 17, 2014

Abstract

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.

Keywords: Public goods game · representative sample · experiments · external validity

JEL classification: C72 · C90 · Z13

I Introduction

The issue of external validity of lab 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 behaviour 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

*Bortolotti: Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy; stefania.bortolotti@unibo.it, Phone: +39051 209 8135.

†Casari (corresponding author): Department of Economics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy; marco.casari@unibo.it, Phone: +39051 209 8662, Fax: +39051 209 8493.

‡Pancotto: University of Modena and Reggio Emilia, Viale Allegri, 9, 40121, Italy; francesca.pancotto@unimore.it, Phone: +39 0522 523 264.

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 seems to suggest that students are less pro-social than other subject pools. For instance, in a prisoners dilemma, students cooperate less than white collar workers (Bigoni et al., 2012) or bicycle messengers (Burks et al., 2009). Similarly, students are less pro-social 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 (Carpenter et al., 2005; Dragone et al., 2013). The gap remains when one compares students and *professionals*, i.e. 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 et al., 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

¹The categories studied include business people and managers Alpert (1967); Cooper (2006); Cooper et al. (1999); Croson and Donohue (2006); Dejong et al. (1988); Fehr and List (2004); prisoners (Block and Gerety, 1995), lay people (Glaser et al., 2005); children (Harbaugh et al., 2001, 2002; Murnighan and Saxon, 1998); finance industry professionals (Alevy et al., 2007; Sade et al., 2006); and public affair officials (Potters and Van Winden, 2000).

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

³The choice is task specific, as contractors would not be the most appropriate sample for studying the man-

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.

Table 1: Studies comparing college students with the general population

	Stratified sample	Same procedures for both samples	Lotteries	Dictator	Ultimatum	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 et al. (2012)	Y	Y(mail)						1		Switzerland
Carpenter et al. (2008)	N*	N		1						US
Gächter et al. (2003)	N*	N					1			Russia
Gächter et al. (2004)	N*	N				1				Russia
Belot et al. (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 et al. (2002)	Y	Interview	1							Denmark
Ermisch et al. (2009)	Y	Interview						1		Britain
Exadaktylos et al. (2013)	N*	Interview		1	1			1		Spain
Bellemare et al. (2008)	N*	Internet		1	1					Netherlands
Bellemare et al. (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 pre-specified 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.

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

agement of a renewable natural resource.

society. Let's consider, for instance, a society made up of young and old citizens. Cooperation and punishment behaviours 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 sub-samples.

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 et al. (2002); Bellemare and Kroger (2007); Falk et al. (2012); 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 sub-samples, our study, together with Falk et al. (2012), restricts participation of the student and representative samples to those subjects resident in a given region.

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

⁴Papers comparing students and non-students 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 et al., 2013).

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 behavioural 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, this results does 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 (Ostrom et al., 1992; Herrmann et al., 2008; Henrich et al., 2010). Here we show experimentally that, even within the same culture, punishment has a beneficial or a detrimental effect on cooperation depending on the sub-sample of the population involved. The remainder of the paper 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’ behav-

ior. 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 vouchers 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

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

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

⁷These 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 1st, 2009 (source Istat: <http://demo.istat.it/pop2009/index1.html>).

⁸Because 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, which may restrict the generalizability of lab results obtained with students: (i) there is an endogenous sample selection among students participating in experiments; (ii) students are not informative about the general population. Since we are mainly interested in (ii), 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 non-students that used to take part in experiments, we however retain the term ‘Student’ for brevity.

⁹Since 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.

¹⁰The database includes a small fraction of non-students, most of them were former students living in the area. We had 18 non-student participants (17 percent). Thanks to a questionnaire we know that 14 are 32 years old

Table 2 reports the socio-demographic characteristics of the two samples. Whereas gender composition is similar, age and employment compositions differ widely. In the *Student* sample the vast majority of participants is aged between 18-39, while in the *Representative* sample most participants belong to the 40-59 category (44.7 percent) 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 percent 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 may suggest shared social norms: as a matter of fact, about 87 (84) percent 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 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:

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

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.

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

¹²Each 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 Appendix.

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

Table 2: Socio-demographic 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		
	02/03/2011	23/02/2011
	04/03/2011	24/03/2011
Dates (dd/mm/yyyy)	05/03/2011	24/03/2011
	01/10/2011	16/06/2012
No. of participants	108	104

Notes: Self-reported answers from a post-experimental computerized questionnaire. Due 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 out of 108 subjects.

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:

$$\pi_i = \pi_i^1 - b \sum_{j \neq i}^N p_j^i - \sum_{j \neq i}^N p_i^j$$

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 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 centre, where we deployed the mobile Bologna Laboratory for Experiments in Social Sciences (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) 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

¹⁴Upon 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).

¹⁵In 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/>).

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, 3) and then present the evolution of contributions and punishment norms over time (Results 4 and 5).

Table 3: Average contributions to the public good

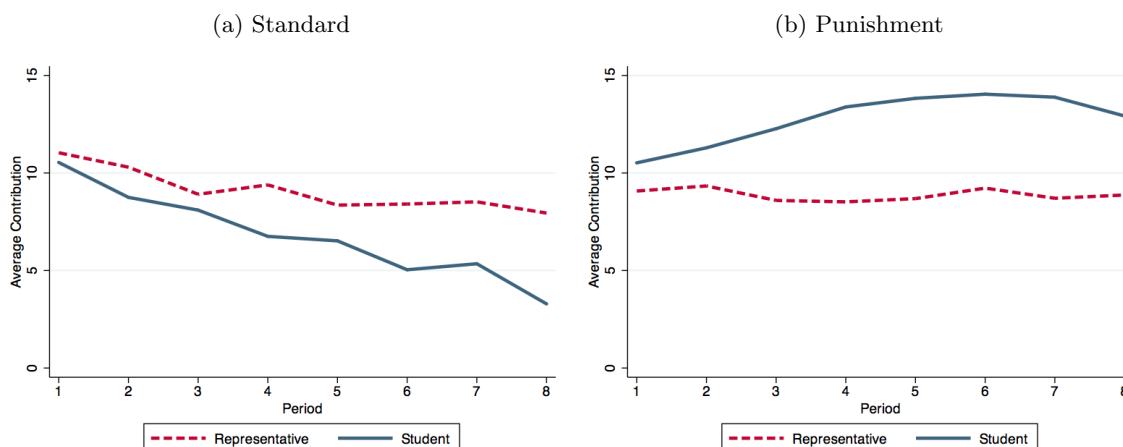
PGG	All periods		First Period	
	Representative	Student	Representative	Student
Standard	9.11	6.79	11.04	10.54
Punishment	8.88	12.77	9.07	10.52

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

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.*

Figure 1: Contributions to the public good over time



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, $p=0.149$, $N_R = N_S = 4$, two-sided).

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.*

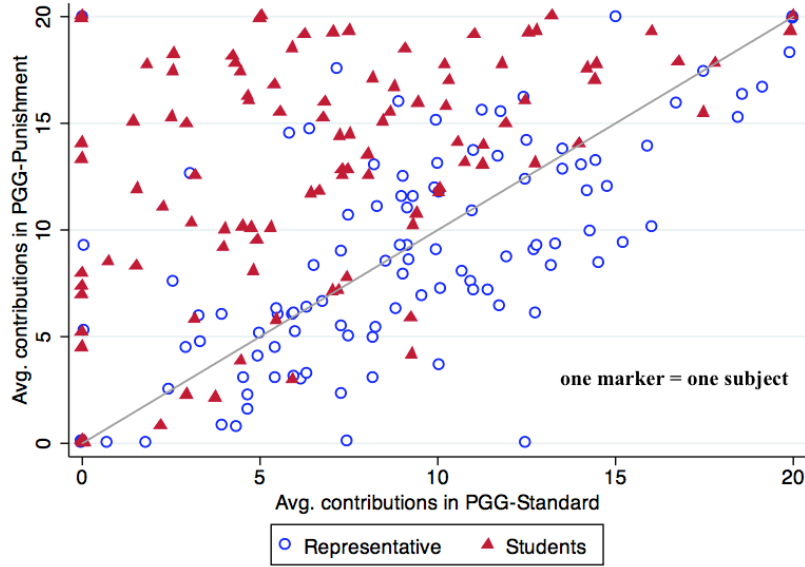
The introduction of peer punishment reverses the treatment order: the general population contributes less as compared to students. Average cooperation in the PGG-Punishment was 8.9 in the *Representative* and 12.8 in the *Student* treatment. Support for Result 2 comes from Table 4 and Figure 2.

The difference across subject pools in the PGG-Punishment is highly significant according

¹⁶We 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.

¹⁷In addition, we control for three alternative ways of modelling low understanding: *i*) subjects in the last quartile of the distribution according to either the number of mistakes or the total answering time; *ii*) subjects without a college degree or higher; *iv*) 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 2: Average individual contributions in PGG-Standard and PGG-Punishment



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 ($p=0.021$, $N_R = N_S = 4$, 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, $p=0.068$, $N_{S-PGG-std} = N_{S-PGG-punish} = 4$, $p=0.465$, $N_{R-PGG-std} = N_{R-PGG-punish} = 4$, 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 percent of students contribute on average more with than without punishment opportunities (vs. 32 percent 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 analyse the initial contribution levels as well as the dynamics of contribution and punishment over time (Figure 1).

Result 3 *Cooperation in the initial period is indistinguishable between representative and stu-*

¹⁸Per-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 4: Treatment effect on contributions.

	Dep.var.: Contribution					
	PGG Standard		PGG Punishment			
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Representative Sample	0.842*** (0.252)	0.676*** (0.261)	-0.117 (0.327)	-1.487*** (0.309)	-1.631*** (0.326)	-0.573 (0.367)
Low understanding	-0.018 (0.337)		-0.019 (0.362)	-0.506 (0.412)		-0.527 (0.421)
4 or more mistakes		0.576** (0.284)			0.500 (0.352)	
Period			-0.365*** (0.033)			0.192*** (0.032)
Period x Representative			0.237*** (0.044)			-0.211*** (0.043)
N.obs.	1696	1696	1696	1696	1696	1696

Notes: Ordered logit regression on individual contribution levels, individual-level random effects. Symbols ***, **, and * indicate significance at the 1%, 5% and 10% level, respectively.

dent 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, $p=0.614$, $N_R = 108$, $N_S = 104$, two-sided). The same conclusion holds for the PGG-Punishment ($p=0.169$, $N_R = 108$, $N_S = 104$, two-sided).¹⁹ We also regressed contributions in the first period over the dummy *Representative sample* (see Table A-1 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 while 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, while 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.

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

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 – i.e., 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 et al., 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). 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 $t-1$ x Representative sample* is indeed neg-

²⁰Conditional 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.

²¹If 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.

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

²³As a robustness check, we run the same regressions using a 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 Appendix.

Table 5: Conditional cooperation and observed contributions.

<i>Dep. var.:</i>	PGG-Standard			PGG-Punishment		
	Representative Model 1	Students Model 2	Pooled sample Model 3	Representative Model 4	Students Model 5	Pooled sample Model 6
<i>Contribution</i>						
Others' contributions in t-1	0.004 (0.006)	0.030*** (0.007)	0.035*** (0.006)	0.021*** (0.007)	0.042*** (0.007)	0.043*** (0.007)
Period	-0.099*** (0.037)	-0.284*** (0.045)	-0.173*** (0.028)	-0.012 (0.036)	0.071* (0.042)	0.022 (0.027)
Low understanding	0.560 (0.475)	-0.520 (0.590)	0.072 (0.370)	-0.284 (0.564)	-0.796 (0.710)	-0.515 (0.446)
Representative Sample			1.760*** (0.356)			-0.693 (0.458)
Others' contr in t-1 x Representative			-0.034*** (0.009)			-0.021** (0.010)
N.obs.	756	728	1484	756	728	1484

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

ative, 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 *pro-social* when the target of the punishment is a free-rider; conversely, we say punishment is *anti-social* 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?

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); while 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 pro-social and less anti-social 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 pro-social vs. anti-social punishment. In line with other studies, pro-social punishment is more frequent than anti-social 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 anti-social punishment.

²⁴A Wilcoxon rank-sum test does not reveal any statistically significant difference when taking each session as an independent observation ($p=0.149$, $N_R = N_S = 4$).

Figure 3: Received punishment by contribution level.

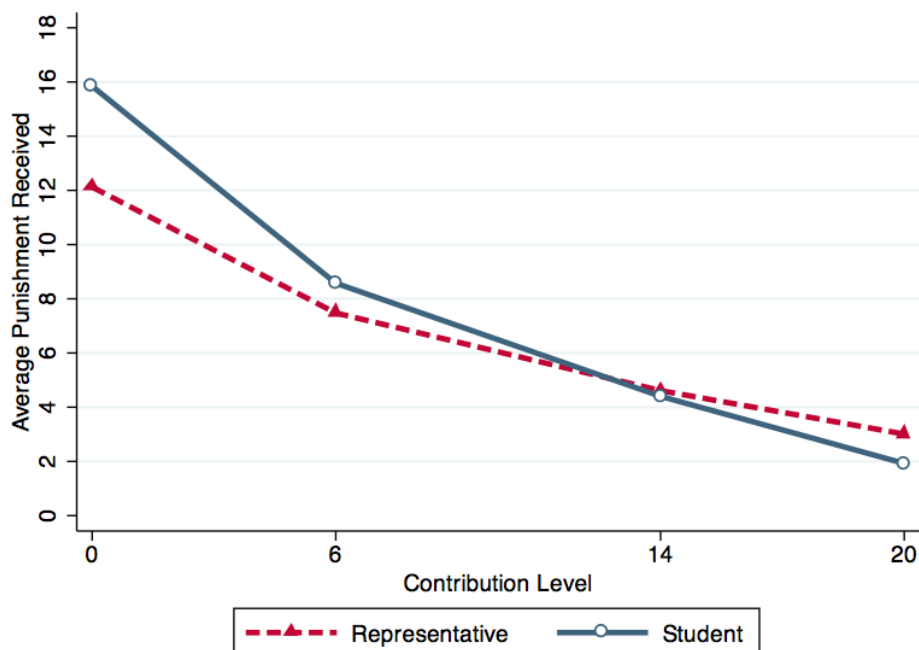


Table 6: Received punishment by contribution level.

	<i>Dep.var.: Deductions assigned (1=Yes; 0=No)</i>			
	$x_i = 0$	$x_i = 6$	$x_i = 14$	$x_i = 20$
	Model 1	Model 2	Model 3	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)
N.obs.	1074	1239	1344	1431

	<i>Dep.var.: Deductions assigned (1=Yes; 0=No)</i>			
	$x_i = 0$	$x_i = 6$	$x_i = 14$	$x_i = 20$
	Model 1	Model 2	Model 3	Model 4
Representative Sample	-1.089** (0.473)	-0.404 (0.371)	-0.234 (0.416)	0.991* (0.545)
4 or more mistakes	-0.541 (0.494)	0.160 (0.397)	0.541 (0.452)	1.179** (0.583)
N.obs.	1074	1239	1344	1431

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

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

Table 7: Treatment effect on punishment.

<i>Dep.var.:</i>	Representative	Students	Pooled sample
<i>Deduction points received</i>	Model 1	Model 2	Model 3
Other's contributions	-0.019*** (0.006)	-0.033*** (0.007)	-0.032*** (0.007)
Positive deviation	0.104*** (0.016)	0.059*** (0.023)	0.062*** (0.022)
Negative deviation(abs)	0.209*** (0.019)	0.356*** (0.022)	0.341*** (0.018)
Period	-0.033 (0.027)	-0.019 (0.031)	-0.027 (0.020)
Representative Sample			0.469 (0.365)
Others' contrib. x Representative			0.012 (0.009)
Pos. deviation x Representative			0.041 (0.027)
Neg. deviation x Representative			-0.125*** (0.024)
N.obs.	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* take values other than 0 when a subject's contribution is larger than the average contribution of the others. Symbols ***, **, and * indicate significance at the 1%, 5% and 10% level, respectively.

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*).

²⁵Models 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.

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 $t-1 \times Representative$* 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 socio-demographic 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 socio-demographic dimensions: college students and a representative sub-sample 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 et al., 2008; Cappelen et al., 2010; Belot et al., 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 et al.,

Table 8: Variation in contribution levels and punishment: high vs. low contributors.

<i>Dep. var.: Delta contributions</i> 1 if $ Give_t - Give_{t-1} > 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 (0.141)	0.453** (0.230)	0.428** (0.213)	0.293 (0.311)	0.806*** (0.204)	0.953*** (0.218)
Period	0.048 (0.103)	0.038 (0.165)	0.039 (0.087)	0.200 (0.143)	-0.009 (0.092)	0.064 (0.078)
Low understanding	0.491 (0.747)	0.085 (1.072)	0.324 (0.612)	-0.261 (1.325)	1.060* (0.623)	0.658 (0.622)
Representative Sample			2.302** (1.112)			2.650*** (0.559)
Deduction received in t-1 x Representative			-0.534** (0.258)			-0.747** (0.333)
N.obs.	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. Symbols ***, **, and * indicate significance at the 1%, 5% and 10% level, respectively.

2008). Here we show that, even within the same society, the impact of peer punishment in promoting cooperation can vary widely depending on the sub-sample of the population considered. Our results document that punishment can reverse the ordering of sub-groups 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 sub-sample, whereas little or no effect was detected in 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., anti-social punishment) and this happened with a higher frequency than in the students 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 sub-sample. 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 sub-sample increase with repetition while they remain flat in the general population.

More generally, a main behavioural difference between the sub-samples 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 sub-samples 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 students 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.

Acknowledgments. 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.

References

- Alevy, J. E., M. S. Haigh, and J. A. List (2007). Information cascades: Evidence from a field experiment with financial market professionals. *The Journal of Finance* 62(1), 151–180.
- Alpert, B. (1967). Non-businessmen as surrogates for businessmen in behavioral experiments. *The Journal of Business* 40(2), 203–207.
- Bellemare, C. and S. Kroger (2007). On representative social capital. *European Economic Review* 51(1), 183 – 202.
- Bellemare, C., S. Kroger, and A. van Soest (2011). Preferences, intentions, and expectation violations: A large-scale experiment with a representative subject pool. *Journal of Economic Behavior & Organization* 78(3), 349 – 365.
- Bellemare, C., S. Krger, and A. Van Soest (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica* 76(4), 815–839.
- Belot, M., R. Duch, and L. Miller (2010). 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.
- Bigoni, M., S. Bortolotti, M. Casari, D. Gambetta, and F. Pancotto (2013). Cooperation hidden frontiers: The behavioral foundations of the italian north-south divide. Technical report, Department of Economics, University of Bologna WP 882.
- Bigoni, M., M. Casari, and G. Camera (2012). Strategies of cooperation and punishment among students and clerical workers. *Journal of Economic Behavior and Organization*.
- Block, M. K. and V. E. Gerety (1995). Some experimental evidence on differences between student and prisoner reactions to monetary penalties and risk. *The Journal of Legal Studies* 24(1), 123–38.
- Bosch-Domenech, A., J. Montalvo, R. Nagel, and A. Satorra (2002). One, two, (three), infinity,...: Newspaper and lab beauty-contest experiments. *American Economic Review* 92(5), 1687–1701.

- Bram Cadsby, C. and E. Maynes (1998, October). Choosing between a socially efficient and free-riding equilibrium: Nurses versus economics and business students. *Journal of Economic Behavior & Organization* 37(2), 183–192.
- Burks, S., J. Carpenter, and L. Goette (2009). Performance pay and worker cooperation: Evidence from an artefactual field experiment. *Journal of Economic Behavior & Organization* 70(3), 458 – 469.
- Butler, D. M. and T. Kousser (2013). How do public goods providers play public goods games?
- Camerer, C. F. (2003). *Behavioral Game Theory*. New York, NY: Russell Sage Foundation.
- Cappelen, A. W., K. Nygaard, E. O. Sørensen, and B. Tungodden (2010). 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.
- Cardenas, J. C. (2005). Groups, commons and regulations: Experiments with villagers and students in Colombia. *Psychology, Rationality and Economic Behaviour*, 242.
- Carpenter, J., C. Connolly, and C. Myers (2008). Altruistic behavior in a representative dictator experiment. *Experimental Economics* 11, 282–298.
- Carpenter, J. and E. Seki (2011). Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. *Economic Inquiry* 49(2), 612–630.
- Carpenter, J. P., S. Burks, and E. Verhoogen (2005). *Comparing students to workers: The effects of social framing on behavior in distribution games*, pp. 261–289. Emerald Group Publishing Limited, pp.
- Cooper, D. J. (2006). Are experienced managers experts at overcoming coordination failure? *The B.E. Journal of Economic Analysis & Policy* 6(2).
- Cooper, D. J., J. K. W. Lo, and Q. L. Gu (1999). Gaming against managers in incentive systems: Experimental results with Chinese students and Chinese managers. *American Economic Review* 89(4), 781–804.
- Croson, R. and K. Donohue (2006). Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science* 52(3), pp. 323–336.

- Dejong, D. V., R. Forsythe, and W. C. Uecker (1988). A note on the use of businessmen as subjects in sealed offer markets. *Journal of Economic Behavior & Organization* 9(1), 87–100.
- Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2008). Representative trust and reciprocity: Prevalence and determinants. *Economic Inquiry* 46(1), 84–90.
- Dragone, D., F. Galeotti, and R. Orsini (2013). Temporary workers are not free-riders: An experimental investigation. Technical report, University of Bologna, DSE-WP 915.
- Dyer, D., J. H. Kagel, and D. Levin (1989). A comparison of naive and experienced bidders in common value offer auctions: A laboratory analysis. *The Economic Journal* 99(394), 108–115.
- Egas, M. and A. Riedl (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B* 275, 871–878.
- Ermisch, J., D. Gambetta, H. Laurie, T. Siedler, and S. C. Noah Uhrig (2009). Measuring people’s trust. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172(4), 749–769.
- Exadaktylos, F., A. Espin, and P. Branas-Garza (2013). Experimental subjects are not different. *Nature, Scientific Reports* (3).
- Falk, A., S. Meier, and C. Zehnder (2012). Do lab experiments misrepresent social preferences? *Journal of the European Economic Association* forthcoming.
- Fehr, E. and S. Gächter (2000). Cooperation and punishment in public goods experiments. *American Economic Review* 90(4), 980–994.
- Fehr, E. and J. A. List (2004). The hidden costs and returns of incentives-trust and trustworthiness among CEO’s. *Journal of the European Economic Association* 2(5), 743–771.
- Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2), 171–178.
- Fischbacher, U., S. Gächter, and E. Fehr (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71(3), 397–404.
- Fréchette, G. R. (2009). Laboratory experiments: Professionals versus students. *The Methods of Modern Experimental Economics*. Oxford University Press, forthcoming.

- Gächter, S., B. Hermann, and C. Thöni (2003). Norms of cooperation among urban and rural dwellers. experimental evidence from russia. *Mimeo, University of St. Gallen.*
- Gächter, S., B. Hermann, and C. Thöni (2004). Trust, voluntary cooperation, and socio-economic background: Survey and experimental evidence. *Journal of Economic Behavior and Organization* 55, 505–531.
- Gächter, S. and B. Herrmann (2011). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review* 55(2), 193–210.
- Glaser, M., T. Langer, and M. Weber (2005). Overconfidence of professionals and lay men: Individual differences within and between tasks? *Sonderforschungsbereich 504, Universität Mannheim & Sonderforschungsbereich 504, University of Mannheim.*
- Greiner, B. (2004). 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.
- Harbaugh, W. T., K. Krause, and T. R. Berry (2001). Garp for kids: On the development of rational choice behavior. *The American Economic Review* 91(5), 1539–1545.
- Harbaugh, W. T., K. Krause, and L. Vesterlund (2002). Risk attitudes of children and adults: Choices over small and large probability gains and losses. *Experimental Economics* 5(1), 53–84.
- Harrison, G. W., M. I. Lau, and M. B. Williams (2002). Estimating individual discount rates in Denmark: A field experiment. *The American Economic Review* 92(5), pp. 1606–1617.
- Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic Literature* 42(4), 1009–1055.
- 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 (2010). Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. *Science* 327(5972), 1480–1484.
- Herrmann, B., C. Thöni, and S. Gächter (2008). Antisocial Punishment Across Societies. *Science* 319(5868), 1362–1367.

- Kocher, M. G., T. Cherry, S. Kroll, R. J. Netzer, and M. Sutter (2008). Conditional cooperation on three continents. *Economics Letters* 101(3), 175 – 178.
- Murnighan, J. K. and M. S. Saxon (1998). Ultimatum bargaining by children and adults. *Journal of Economic Psychology* 19(4), 415–445.
- Ostrom, E., J. Walker, and R. Gardner (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review* 86, 404–417.
- Potters, J. and F. Van Winden (2000). Professionals and students in a lobbying experiment: Professional rules of conduct and subject surrogacy. *Journal of Economic Behavior & Organization* 43(4), 499–522.
- Sade, O., C. Schnitzlein, and J. F. Zender (2006). Competition and cooperation in divisible good auctions: An experimental examination. *Review of Financial Studies* 19(1), 195–235.