University of Heidelberg

Department of Economics

Discussion Paper Series | No. 595

What do we learn from public good games about voluntary climate action? Evidence from an artefactual field experiment

Timo Goeschl, Sara Elisa Kettner,
Johannes Lohse, and Christiane Schwieren

June 2015

What do we learn from public good games about voluntary climate action? Evidence from an artefactual field experiment *

Timo Goeschl[†] Sara Elisa Kettner[‡] Johannes Lohse[§]
Christiane Schwieren[¶]

June 16, 2015

Abstract

Evidence from public good game experiments holds the promise of instructive and cost-effective insights to inform environmental policy-making, for example on climate change mitigation. To fulfill the promise, such evidence needs to demonstrate generalizability to the specific policy context. This paper examines *whether* and *under which conditions* such evidence generalizes to voluntary mitigation decisions. We observe each participant in two different decision tasks: a real giving task in which contributions are used to directly reduce CO_2 emissions and a public good game. Through two treatment variations, we explore two potential shifters of generalizability in a within-subjects design: the structural resemblance of contribution incentives between the tasks and the role of the subject pool, students and non-students. Our findings suggest that cooperation in public good games is linked to voluntary mitigation behavior, albeit not in a uniform way. For a standard set of parameters, behavior in both tasks is uncorrelated. Greater structural resemblance of the public goods game leads to sizable correlations, especially for student subjects.

Keywords: Public Goods; Experiments; Climate Change

*The authors gratefully acknowledge financial support by the German Ministry for Education and Research under grant 01UV1012. Furthermore we would like to thank the audiences at ESA New York, ESA Cologne, AURÖ Bern, ZEW Mannheim, WCERE Istanbul, and the IfW Kiel for their valuable comments.

[†]Email: goeschl@eco.uni-heidelberg.de. Postal address: Heidelberg University, Department of Economics, Bergheimer Str. 20, 69115 Heidelberg, Germany.

[‡]Email: kettner@eco.uni-heidelberg.de. Postal address: Heidelberg University, Department of Economics, Bergheimer Str. 58, 69115 Heidelberg, Germany

[§]Email: lohse@eco.uni-heidelberg.de. Postal address: Heidelberg University, Department of Economics, Bergheimer Str. 20, 69115 Heidelberg, Germany. Phone: +49 6221 548013

[¶]Email: christiane.schwieren@awi.uni-heidelberg.de. Postal address: Heidelberg University, Department of Economics, Bergheimer Str. 58, 69115 Heidelberg, Germany

1 Introduction

Economists typically treat climate change mitigation as a public goods problem (Nordhaus, 1991). Consequently, most theoretical models of voluntary mitigation efforts predict that free-riding is the dominant individual behavior. Empirically, however, public good game (PGG) experiments and other social preference tasks have amassed convergent evidence that free-riding may be less prevalent in social dilemmas than predicted (Ledyard, 1995; Zelmer, 2003; Chaudhuri, 2011; Vesterlund, 2014). Does this experimental evidence give reason to rethink the premises of climate policies that are designed with large-scale free-riding in mind? And more generally, can PGG and variants thereof serve as a reliable testbed for predicting behavioral responses to climate change policies?

A number of recent papers tentatively argue that findings from PGG experiments could provide important insights into mitigation behavior and policies in the real world (Shogren and Taylor, 2008; Venkatachalam, 2008; Brekke and Johansson-Stenman, 2008; Gowdy, 2008; Gsottbauer and van den Bergh, 2011; Carlsson and Johansson-Stenman, 2012). In the same spirit, some experimental studies on public good provision have been framed or interpreted with an explicit reference to mitigation decisions (e.g., Milinski et al., 2006, 2008; Tavoni et al., 2011; Brick and Visser, 2015). Such experiments present a theoretically appealing method for obtaining causal evidence at low cost. However, whether or not such PGG experiments can truly provide the desired valuable insights crucially depends on their *generalizability* (Levitt and List, 2007), i.e., the degree to which generic behavior, based on observing subjects in an abstract lab task, transfers to the specific context of mitigation decisions. Whether and under which conditions behavior in a PGG experiment generalizes to voluntary mitigation choices is, at heart, an empirical question. In the present paper, we take a first step towards providing an answer.

Concerns that subjects' behavior in abstract game forms under controlled conditions in the laboratory may not generalize to individual behavior in context-rich situations outside the lab are not new. But their recent recurrence in the context of whether social preferences elicited using standard experimental designs are predictive beyond the lab (Levitt and List, 2007), is particularly relevant for issues of public goods provision such as voluntary mitigation choices.[1] Evidence on generalizability in this context is mixed: The extent to which cooperation in PGG correlates with a broader set of pro-social preferences (Blanco et al., 2011; Peysakhovich et al., 2014) and, more importantly, the extent to which it generalizes to cooperative behavior beyond the laboratory (Benz and Meier, 2008; Laury and Taylor, 2008; de Oliveira et al., 2011; Voors et al., 2012) varies substantially across studies. On the basis of available evidence, generalizability of behavior in the PGG to voluntary mitigation choices can therefore neither be ruled in nor out.

The climate context, to which one hopes to generalize PGG evidence, provides additional reasons for concern. It could be argued that the deliberately abstract format of the PGG does not capture context-specific preferences (e.g., risk- or time-preferences), beliefs (e.g., regard-

[1] Levitt and List(2007) describe a number of situational factors, present in a typical lab experiment, that might reduce its predictive power for field behavior. For instance, they discuss the extent of scrutiny, the activation of specific norms, or the context in which the decision is embedded as important shift parameters. Their concerns, arguably, carry more weight for experiments conducted in order to inform policy makers than for experiments that try to falsify a theory (Schram, 2005; Sturm and Weimann, 2006; Kessler and Vesterlund, 2015).

ing the expected damages from climate change), or attitudes (e.g., regarding the importance of pro-environmental behavior) that, at least in theory, should also shape voluntary mitigation decisions. On the other hand, the experimental paradigm of the PGG can accommodate considerable variation in design features. For instance, a greater resemblance to voluntary mitigation decisions might result from simple changes to design parameters such as the group size or the productivity of the experimental public good. If such variations are able to capture most of the relevant drivers of mitigation decisions, then generalizability may be accomplishable at acceptable cost.

This paper brings new experimental evidence to two of the issues raised above. First, we examine whether estimates of generic cooperative preferences derived from behavior in a PGG experiment can explain a significant portion of individual mitigation behavior, as opposed to unobserved idiosyncratic motives. Such explanatory power of sufficient size is an important prerequisite for a high level of generalizability (Al-Ubaydli and List, 2015). The empirical problem is that the totality of individual mitigation behavior, just like the totality of an individual's charitable behavior towards others, is not observable for the researcher.[2] Following other examples in the literature (Benz and Meier, 2008; Laury and Taylor, 2008; de Oliveira et al., 2011; Voors et al., 2012), we approximate the ideal test by conducting a laboratory experiment in which we observe each participant in two contribution situations: A public goods game and a real giving task in which contributions are used to reduce CO_2 emissions.

The second issue that we examine within this framework is whether abstract PGG experiments can be implemented in a way that increases the generalizability of its output in the direction of voluntary mitigation choices. We do so by experimentally varying two central design features of how PGG evidence is generated, namely its parameter structure and the subject pool. The systematic variation of PGG parameters, in particular group size, marginal per-capita return (MPCR), and payoff symmetry, allows us to test whether generalizability varies with different degrees of structural resemblance between PGG contribution incentives and voluntary mitigation incentives. The comparison of behavior across two samples, one a sample of students and one recruited from the general population, allows us to test whether generalizability in a climate context perhaps hinges on subject pool. It is well know that student samples, which account for the majority of PGG evidence, share only a limited range of individual attributes with the general population. As a result, the extent to which the behavior of the former allows conclusions about the latter is a matter of ongoing discussion (Gächter et al., 2004; List, 2004; Carpenter et al., 2008; Thöni et al., 2012; Anderson et al., 2013; Falk et al., 2013; Belot et al., 2015) and at the same time the source of uncertainty over its generalizability to mitigation actions.

Our results suggest that PGG behavior can be indicative of voluntary mitigation decisions, but not in a uniform fashion. Instead, the potential for generalizability crucially depends on the way the PGG is designed and conducted. In a benchmark case employing common PGG parameters, the correlation between contributions in both task is small and insignificant. This result holds

[2] Under ideal conditions, the researcher would observe two separate decisions by the same individual: Contribution choices in a standard PGG and revealed preferences for voluntary CO_2 mitigation in a field context. The latter would require observing the totality of economic decisions that potentially involve a direct or indirect mitigation of CO_2 emissions. In a fossil-fuel economy, this is true for almost all economic decisions. Accurate measurement of the aggregate pure mitigation effort at the level of the individual is therefore empirically daunting, particularly if this measurement should also be obtained in an unintrusive fashion

irrespective of the subject pool. A low correlation indicates that there exist idiosyncratic drivers of mitigation behavior that remain unobserved in standard PGG. Yet, when PGG parameters resemble more closely the incentive structure underlying voluntary climate change mitigation, correlations - especially those for student subjects - become significant and sometimes sizable. Thus, by implementing simple design changes, some of the apparent differences in individual behavior disappear. This points towards a cost-effective and feasible way of improving current insights into the institutional mechanisms affecting voluntary mitigation behavior that can be gained via laboratory experiments. On the other hand, switching to a subject pool of non-students has more ambiguous effects. In line with previous results, we find that on average, non-students contribute more in both tasks. However, as indicated by strongly reduced correlations, the degree of generalizability is much lower within this more heterogeneous sample. This underlines the existence of a trade-off between representativeness and generalizability unless the apparatus of the experimental design or the sample size are significantly enlarged - at a cost.
The remainder of the paper is organized in the following way: Section 2 discusses our research question in relation to the existing literature. In Section 3, we describe the experimental set-up and the characteristics of our subject pool. Section 4 contains the analyses and core results. Section 5 concludes with a discussion of our findings.

2 Related Literature

There are several studies that examine issues of generalizability, both regarding the relationship of social preferences measured in different abstract lab tasks (e.g., Public Good Game; Dictator Game; Trust Game) and regarding the predictive power of cooperative behavior observed in PGG towards contributions made to a variety of naturally occurring public goods. We follow these studies in their common methodology of employing a within-subjects design.
So far only few studies have analyzed how cooperation in public good games corresponds to social preferences elicited in other abstract tasks. Overall, these studies arrive at mixed results. Blanco et al. (2011) find that contributions made in a standard PGG are significantly correlated with responders' behavior in a sequential prisoners dilemma, but not to other-regarding choices made in ultimatum or dictator games. In an online experiment, Peysakhovich et al. (2014) find stronger evidence that an individual's propensity to contribute in a one-shot public good game spills over to other abstract game formats. More cooperative subjects are shown to be significantly more likely to give higher amounts in a dictator game and to reciprocate trusting behavior more strongly in a trust game. They furthermore find that more cooperative subjects are also more prone to help the experimenters after the actual experiment, by voluntarily completing an additional questionnaire. Finally, in Galizzi and Navarro Martinez (2015) public good game behavior is moderately, but significantly correlated with behavior in trust and dictator games.[3] This first strand of literature highlights that the same individual can behave quite differently even in related abstract social preference tasks, in which idiosyncratic motives should be largely absent.

[3]They, however, detect no significant relationship with helping or donation behavior in five different field situations which are randomly administered subsequent to the actual experimental sessions.

A second strand of literature addresses the same basic question as our paper by investigating the relationship between contributions observed in a laboratory public goods game and contributions to a naturally occurring public good. As in our experiment, these studies largely lack a direct and unintrusive measure of cooperation in the field.[4] Instead, they observe contributions to a naturally occurring public good through eliciting choices in a modified dictator game (Eckel and Grossman, 1996). Benz and Meier (2008) investigate the correlation between students' charitable giving in a laboratory setting and their charitable giving in an university fund-raiser. Within a low-income neighborhood, de Oliveira et al. (2011) explore whether subjects who display other-regarding preferences in a linear public goods game also give to local charities. Voors et al. (2012) compare the behavior of subsistence farmers in a linear public goods game to the amount they contribute to a real community public good. Closest to our own question, Laury and Taylor (2008) investigate student behavior in a variety of the linear public good game and their contributions to a local environmental public good. These studies have brought forth mixed results: some of them find a significant correlation between contributions in the abstract and specific context (Benz and Meier, 2008), whereas others suggest a more moderate (Laury and Taylor, 2008; de Oliveira et al., 2011) or even insignificant (Voors et al., 2012) relationship. In a comprehensive literature review, Galizzi and Navarro Martinez (2015) similarly conclude that results vary greatly across studies according to their context (e.g., the real public good offered) and design (e.g., the subject pool under study or the experimental procedures used to assess generic cooperation rates).

In light of the literature reviewed above, the extent to which existing findings are transferable to the specific context of voluntary climate change mitigation is not clear. Several design differences plausibly limit transferability: First, all of the studies above use a particular local public good, while climate change mitigation is a global and intergenerational public good. Second, each of these four studies was conducted with a specific subject pool of either students or aid recipients. This puts into question whether they are sufficiently representative for reaching conclusions about the behavior of broader segments of the population relevant in a climate policy context. Third, each of these studies - with the exception of Laury and Taylor (2008) - uses one specific set of parameters when assessing generic preferences for cooperation within a PGG.

These plausible limitations to transferability inform important design choices in our experiment, with a view to answering the questions raised in the introduction. Our design employs a task directly linked to the reduction of CO_2 emissions. Furthermore, we use an unified design in which we observe behavior of two different subject pools: One convenience sample of students and a group of subjects that more closely covers demographic attributes of everyday decision-makers. Finally, our design identifies to what degree the correlation between the two tasks depends on the parameter choice in the PGG. These design elements are well suited to provide answers to our research questions with their focus on generalizability to voluntary mitigation.[5]

[4] A notable exception is Fehr and Leibbrandt (2011), in which the overexploitation of a fishery resource is related to behavior in a public good experiment.

[5] Note, however, that the design is explicitly not intended to resolve the broader controversy (Levitt and List, 2007, 2009; Falk and Heckman, 2009; Kessler and Vesterlund, 2015; Camerer, 2015) on whether social-preferences assessed in abstract lab tasks are generally externally valid, in any chosen context.

3 Experimental design and implementation

Questions of generalizability from one experimental task to another are typically addressed by a within-subjects design (Benz and Meier, 2008; Laury and Taylor, 2008; de Oliveira et al., 2011; Blanco et al., 2011; Voors et al., 2012; Peysakhovich et al., 2014). Therefore, we observe for each subject choices in a context-free decision task and in a task related to climate change mitigation. Participants are informed in the initial instructions that there would be several consecutive tasks in which they could earn real money. In *Task I* we assess individual contributions to the real public good of climate change mitigation. In the subsequent *Task II*, subjects take ten one-shot public good decisions in which we vary experimental parameters along three dimensions (Goeree et al., 2002).[6] In the following, we describe each of the decision tasks in more detail.

3.1 Task I: The real contribution task

To observe contributions to climate change mitigation in a lab setting, we employ a real giving task (Eckel and Grossman, 1996) in which individual contributions are used to reduce global CO_2 emissions. The transparent and verifiable reduction is executed by retiring emission permits from the EU ETS (Löschel et al., 2013; Diederich and Goeschl, 2014).[7] Prior to reaching the first decision screen, subjects were informed that they had received 10€ as a reward for taking part in the experiment. Subsequently, they were given the choice to contribute any share of these 10€ (in steps of 1€) towards a common account that would be used by the experimenters to reduce global CO_2 emissions.

Before subjects could select their preferred contribution level on the decision screen, they received a short and neutral description of the public good on an information screen. Thereby we ensured that each subject would have at least the same level of information about greenhouse gas emissions and the procedure by which the emission reductions would be executed by the experimenters. They were also informed about the amount of CO_2 that could be reduced for each 1€-contribution. In order to render the choice tangible, the instructions related this amount to every-day consumption decisions, expressed in terms of two common activities (car travel; use of personal computer) and the average CO_2 emissions of a German citizen. The instructions also confirmed the public good character of CO_2 mitigation by explaining that the particular location of CO_2 reductions would not affect the mitigation of global climate change and by pointing out the temporal delay between the reduction of CO_2 emissions in the atmosphere and the resulting beneficial impacts on climate change.

To avoid potential anchoring effects we made sure that no examples of provision levels were given to subjects before they could select their own contribution. Lastly, participants were informed that documentation from the German Emission Trading Registry would be publicly posted immediately following the last experimental session that would certify that their contributions had been used for the verified emissions reductions.

[6] All subjects in the experiment completed the two tasks in this order. We do not explicitly account for order effects, as Laury and Taylor (2008) find no evidence for such effects in a setting comparable to ours. Furthermore, in a small scale pilot of our study (N=30) we find no evidence for order effects.

[7] Obviously, outcomes from Task I are only a proxy for actual field behavior. But they seem to capture, at least to some degree, environmental preferences, since they are significantly correlated with stated donations to environmental organizations.

3.2 Task II: The laboratory public goods game

The average rate of cooperation in PGG has been found to be responsive to changes in experimental parameters such as the group size, the marginal per capita return (MPCR), or the symmetry of payoffs (Isaac and Walker, 1988; Goeree et al., 2002; Nosenzo et al., 2015). We hypothesize on this basis that the choice of these parameters affects the degree of generalizability. To test this proposition, we employed a variant of the standard public goods game (Goeree et al., 2002): Subjects were anonymously and randomly matched into groups of varying size and completed ten independent one-shot contribution decisions without feedback, displayed on one common decision screen.[8] In each of these decisions participants had to choose how many tokens from their initial endowment they wanted to invest into a public account. Depending on the total number of tokens invested, every public account produced payoffs determined by a distinct combination of MPCR, group size, and payoff symmetry. Table 1 summarizes the ten decisions. In the 'benchmark' or 'reference' case (Decision f), we set the parameters to those used in most existing public good experiments: The group of participants is small, with three members, the payoff structure for investments in the experimental public good is symmetric across participants, and the MPCR is 0.4. In the nine other decisions, the parameter constellation systematically shifted the contribution incentives such that they structurally resembled, to greater or lesser degree, those present in voluntary mitigation decisions. In contrast to the benchmark case, the contribution incentives there are characterized by the fact that the 'group of players' is large, the MPCR is small, and payoffs are asymmetric.

The general payoff structure for individual i is summarized by the following expression:

$$\pi_{it} = v(\omega - x_{it}) + m_t^{int} x_{it} + m_t^{ext} \sum_{j}^{N_t - 1} x_{jt}; \forall i = 1, ..., 12/15; \forall t = 1, ..., 10 \tag{1}$$

where v is the value of a token kept and ω is the initial endowment of tokens. t is a subscript denoting each decision and x_{it} is individuals i's contribution to the public account. m_t^{int} and m_t^{ext} are the internal and external value of a token invested in the public account, respectively. For each token subjects invest in the public account they receive m_t^{int} and transfer m_t^{ext} to every other group member. Cases where $m_t^{int} = m_t^{ext}$ are therefore equivalent to a linear PGG with symmetric payoffs. N_t denotes the number of subjects within a group.

In each decision, tokens remaining in the private account yielded a payoff of $v = 20$ Eurocent and subjects were initially endowed with 20 tokens. As the internal returns are always smaller than v, free-riding is a dominant individual strategy. From the group's perspective, it is efficient to contribute the full endowment. Decisions a-d feature parameters that structurally resemble those for voluntary mitigation decisions (small MPCR, larger group size, and asymmetric payoffs) more than those of the benchmark decision f and decisions g-j.[9]

[8] This screen also contained two additional decisions, not analyzed in this paper. These decisions only served as a robustness check, as they used parameters for which there was no conflict between individual and group interest, and hence, did not resemble a standard public goods problem.

[9] The emphasis here is on structural resemblance. Numerically, of course, the largest feasible group size in a typical lab experiment is still much smaller than the number of beneficiaries of climate change mitigation. The largest group we observe consists of all participants present in a given session, which were either 12 or 15. As a consequence, the lowest MPCR feasible under this constraint is, arguably, still far higher than the potential

Table 1: Parameterization of the 10 PGG Decisions

Decision	Group Size (N)	Internal Return (m_t^{int})	External Return (m_t^{ext})	MPCR	Symmetry
a	12/15	2	2	0.10	Symmetric
b	12/15	3	2	0.10	Advantageous Asymmetric
c	12/15	2	3	0.15	Disadvantageous Asymmetric
d	12/15	4	4	0.20	Symmetric
e	3	8	6	0.33	Advantageous Asymmetric
f	3	8	8	0.4	Symmetric
g	12/15	2	9	0.42	Disadvantageous Asymmetric
h	3	12	8	0.46	Advantageous Asymmetric
i	3	8	12	0.53	Disadvantageous Asymmetric
j	3	16	16	0.80	Symmetric

Notes: This table shows the parameters used in decisions a-j. Internal and external returns are displayed as Eurocent per token contributed to the public account. Decision f is used and marked as reference case, as it is characterized by a combination of parameters that is common in most public good experiments. The MPCR for each decision is calculated by the following formula: $\frac{1}{Nv}(m_t^{int} + (N-1)m_t^{ext})$

To minimize potential bias due to confusion (Houser and Kurzban, 2002; Ferraro and Vossler, 2010), subjects had to go through hypothetical payoff calculations for themselves and other group members, prior to entering the decision screen. In these calculations, there was no pre-specified contribution level to avoid setting a standard. At the end of the experiment, one decision was picked randomly with equal probabilities and payed out to the participants. This randomization of payoffs (Starmer and Sugden, 1991) has the advantage that subjects cannot condition their behavior in a given decision on their other choices.

3.3 Recruitment and sample characteristics

Participant were recruited from two distinct pools. We compare students to non-students in order to analyze, whether the prior focus on student subjects influences the conclusions that can be drawn from existing experiments. To recruit from the general population, we used advertisements in two different local newspapers.[10] As a further recruitment tool, notices about the experiment were posted in all neighborhoods and public places of the city of Heidelberg. Prospective participants contacted a research assistant for further information and were invited to a session.[11] The student sample was recruited from the standard subject pool using ORSEE (Greiner, 2015). To keep the two distinct subject pools comparable in terms of their experience with economic experiments, only subjects who had not taken part in previous studies were included in the experiment. Naturally, both subject pools consist of self-selected subjects. While this is standard practice in almost all economic experiments, there are some concerns that the use of self-selected subjects could overestimate the prevalence of other-regarding preferences (Levitt and List, 2007). Empirically, these concerns have not been confirmed, so far (Anderson et al., 2013; Exadaktylos et al., 2013).

MPCR from avoiding 1 Ton of CO_2.

[10] The *"Rhein-Neckar-Zeitung"* is sold at a price of 1,40 € and has a daily readership of 88.649 within the Heidelberg region. The "Wochen-Kurier" is distributed for free to all households in the Heidelberg region with a run of 74.000 copies.

[11] The research assistant assured that subjects would be able to use a computer. The response rate to the different recruitment methods was comparable and no significant differences can be found with respect to demographic attributes or behavior.

Overall, we recruit 135 subjects for the experiment: 92 from the general population and 43 from the student population. Table 2 gives an overview over the demographic attributes used in the analyses below. The two samples differ significantly with respect to socio-demographics directly related to the student status such as age, income, assets, or number of children. Apart from that, the two pools do not differ significantly regarding their education, stated risk aversion, or stated concern about the consequences of climate change. Obviously, despite being more diverse, the non-student participants in our study are also a convenience sample, but one with a somewhat higher resemblance to the average population.

Table 2: Demographic attributes of different subpopulations

Demographics	**Total**	**Student**	**Non-Student**
	N=135	N=43	N=92
Age (Years)	40.91 (18.76)	22.83 (3.01)	49.36 (16.96)
Gender (1=male)	0.37	0.41	0.35
Individual Net Income (Euro)	1050.83 (902.74)	613.15 (228.59)	1253.65 (1020.73)
Assets (1=Yes)	0.25	0.02	0.36
Education (Years)	14.22 (2.67)	13.86 (1.95)	14.40 (2.94)
Household Size (#)	2.02 (1.44)	1.85 (1.22)	2.10 (1.54)
Has Children (1 = yes)	0.39	0.09	0.53
Stated Risk Aversion (Scale 1 - 11)	4.31 (2.72)	4.27 (2.72)	4.32 (2.73)
Concern Climate Change (Scale 1-7)	5.13 (1.77)	5.04 (1.57)	5.17 (1.87)

Notes: Income is self reported. Assets are coded as a dummy variable that takes the value of one if subjects state that they own either a flat or a house. Risk aversion is self reported based on a question adapted from the German social survey (G-SOEP) ("How do you see yourself: are you in general a person fully prepared to take risks or do you try to avoid taking risks?"). Concerns about climate change are assessed by a questionnaire item ("On a scale of 1-7: How concerned are you about the consequences of climate change")

3.4 Experimental procedures

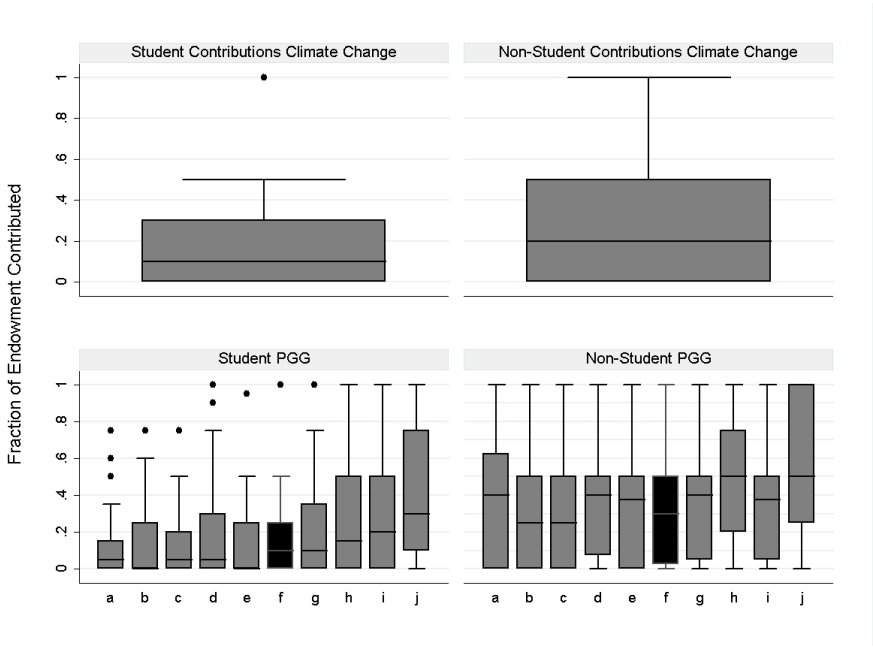
All ten sessions took place at the Heidelberg University Economics Computer Lab using z-Tree (Fischbacher, 2007). There were either 12 or 15 participants per session. At the beginning of a session, participants were seated at one of the available computer terminals, separated by a divider. A printed version of instructions explaining general procedures was handed out and read to subjects before they could begin with the actual decision tasks. All other instructions were fully computerized. Communication between participants was not allowed at any point of the experiment, while questions addressed at the experimenter were answered quietly. All sessions were conducted under full anonymity. Furthermore, communication before the experiment was held at a minimum due to a separate check-in room that reduced common waiting times. In the check-in room subjects also generated a personal code. They were informed up-front that this personal code had the purpose to guarantee their anonymity during the experiment and anonymous payment at the end of a session: Experimenters provided sealed envelopes with earning receipts, only distinguishable by the subjects' personal code. The payment itself was conducted in a different room by a research assistant who was not present at any time of the

experimental sessions. With this payment procedure subjects could be assured that their overall earnings and identity would not be revealed to the experimenter at the end of the session. Sessions lasted around 75 minutes. Average payment was 17.65€ and ranged from 2.68€ to 26.00€ [12].

4 Results

4.1 Observed behavior

Figure 1: Box-plots of contributions across tasks and subject pools



Notes: The top row shows the fraction of endowment contributed to climate change mitigation in the real giving task. The bottom row displays for each decision in the PGG the fraction of endowment contributed to the public account. The black line indicates median contributions. The lower and upper quartiles are marked by the gray box and whiskers are used to display values within 1.5 times the interquartile range. Outliers from this range are displayed as a dot.

Figure 1 gives a first overview over the distribution of contributions in Task I and Task II. The box-plots in the top panel show the fraction of the initial endowment contributed to climate change mitigation during Task I separately for the two different subject pools. The two diagrams in the bottom panel contain information on contribution behavior in Task II. Each box summarizes data for one of the ten distinct public good decisions. In the left diagram we show data for student subjects and in the right one data for non-students. The benchmark case (Decision f) is depicted in a different color.

Median and mean contributions are positive in both tasks and for most parameters values in Task

[12] This value includes earnings from incentivised follow-up questions that are not part of the analysis.

II, contributions in Task I and Task II fall into a similar range.[13] Overall, average contributions in Task I are slightly lower than in Task II, especially for high MPCR decisions.
In line with previous findings (Gächter et al., 2004; List, 2004; Carpenter et al., 2008; Thöni et al., 2012; Anderson et al., 2013; Falk et al., 2013; Belot et al., 2015), student subjects contribute a lower fraction of their initial endowment. Both for the abstract public good decisions in Task II (Mann-Whitney Rank-Sum Test: $p < 0.05$ for each decision) and contributions to climate change mitigation in Task I (Mann-Whitney Rank-Sum Test: $p < 0.05$) this difference is statistically significant. Furthermore, in both tasks a more compressed interquartile range suggest that students' contributions are less dispersed. This observation is also supported by significance tests, which reject the hypothesis of equal variances both for average contributions in Task II (Levene's Robust Test; $p < 0.05$) and contributions in Task I (Levene's Robust Test; $p < 0.001$).
In Task II, the contribution average varies substantially across decisions a-j. In line with previous findings, contributions increase with rising returns from the public good (Goeree et al., 2002). This positive relationship is more pronounced for students than for non-students. Regression results[14] confirm that the fraction of endowment contributed increases significantly with group size ($\beta_1 = 0.021$; $p < 0.001$) and internal ($\beta_2 = 0.030$; $p < 0.001$) or external returns ($\beta_3 = 0.013$; $p < 0.001$). The observation that behavior in Task II depends on the choice of parameters provides a first indication that this design choice could also influence the degree of generalizability from one task to another.

4.2 Individual Behavior: The role of experimental parameters

In this section we study behavior at the individual level to analyze whether and under which conditions PGG experiments capture the main motivational drivers underlying voluntarily carbon emissions reductions, as observed in the real giving task. We answer these two related questions by successively exploring the within-subjects relationship between behavior in Task I and Task II at different levels of aggregation across individuals and Task II decisions. At each of these levels, a high correlation would suggest that contextual factors play a negligible role and behavior in both tasks is driven by generic preferences that favor cooperation.

Result 1: ***There is no significant correlation between average contributions in the abstract public good game and contributions to the real public good of climate change mitigation.***

For a simplified first analysis of the relationship between the two tasks, we follow Laury and Taylor (2008) and initially ignore the variation of parameters between the different decisions of Task II. To broadly summarize contribution behavior, we calculate the mean over the ten distinct public good decisions ($\frac{1}{T}\sum_t^{T=10} x_{it}$). Across all decisions, the average participant contributed 33.85 percent (Median: 32 percent) of his initial endowment to the public account. This average value is close to the cooperation rate (29 percent) reported in Laury and Taylor (2008), who

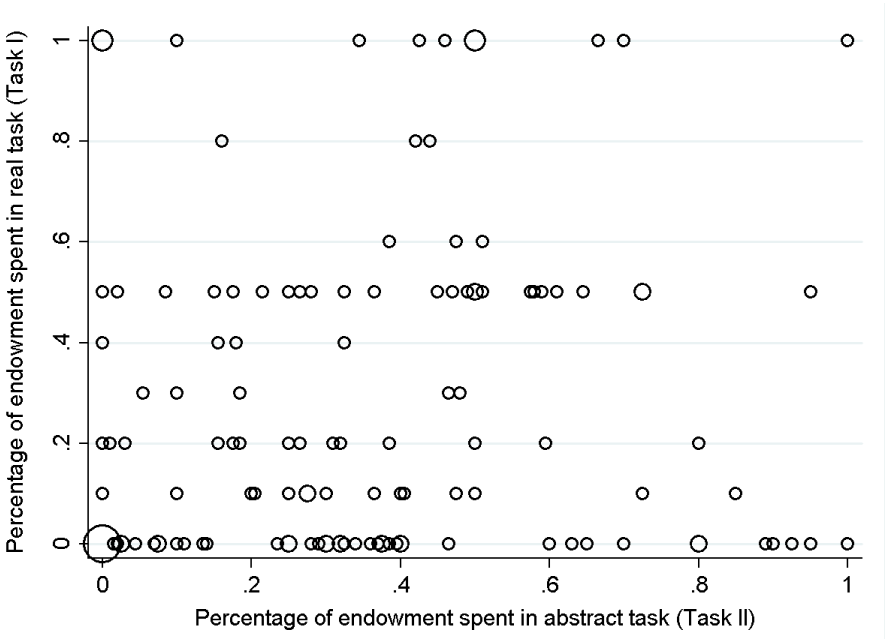
[13] This observation is also supported by non-parametric significance tests (Sign Rank Test: $p < 0.05$) that find significant differences between the tasks for only two out of ten decisions.

[14] We estimate a random effects tobit model controlling for the student status and the set of demographic attributes listed in table 2. Full results are shown in the Appendix table 8.

use a similar PGG design. In comparison, average contributions to climate change mitigation in Task I are only slightly lower at 27.48 percent (Median: 10 percent).

Similar average behavior across tasks need not reflect similar individual behavior. This is, in fact, the main message of figure 2. It shows a bubble plot of realized choices, with the percentage of endowment spent by each individual across all decisions in Task II on the x-axis and that spent in Task I on the y-axis. Visual inspection of the bubble plot does not hint at an association between the size of contributions in the two tasks. The same conclusion arises when employing a relative instead of the absolute scale of contributions: For no more than a quarter of participants do contributions fall into the same quintile in both tasks. The largest overlap can be found within the bottom quintile, a result mostly driven by consistent free-riders. The descriptive results are corroborated by the small and insignificant correlation between contributions in Task I and average contributions in Task II ($r = 0.1303$; $p = 0.132$). In contrast to Laury and Taylor (2008), therefore, behavior in the two distinct tasks in our experiment is only loosely related when the analysis relies on the average decision in Task II.

Figure 2: Scatterplot of average contributions in the PGG and real giving task.



*Notes:*Bubble plot with frequency weights. The size of the bubbles is proportional to the frequency of a pair of contribution choices.

Result 2: ***Correlations are higher when the MPCR in Task II is low, groupsize is large, or payoffs are asymmetric.***

We now move on to explore the correlation structure at a lower level of aggregation of Task II decisions. Thereby we aim to assess how changes in the incentive structure across the ten PGG decisions affect the correlation between contributions made in Task II and Task I. For each decision, table 3 displays the corresponding correlation coefficients for the pooled sample of students and non-students.

Our analysis proceeds in two steps. We first examine the results for decision f. By the choice of parameters (Columns 1-3), this benchmark case is representative for standard public good games. Therefore, decision f is most informative regarding the question to what degree findings from the existing PGG literature readily transfer to the context of climate change. Comparing Task I and decision f of Task II, we find that behavior in the two tasks is not significantly correlated ($r = 0.1404; p = 0.1043$). This cautions against immediate transferability from PGG results to the climate policy context.

As a second step, we turn to the nine other decisions of Task II. Table 3 reports on the correlations. We now see that the relationship between contributions in Task I and Task II strengthens slightly for those Task II decisions that structurally resemble voluntary mitigation decisions: When the MPCR is lower and groups larger than in the benchmark case, we find contribution behavior that is significantly correlated across tasks. The highest significant correlation is reported for decision c, in which there was a low MPCR, a high group size, and an asymmetry of payoffs.[15] Conversely, for those decisions in which the MPCR increases relative to the benchmark case, correlation coefficients drop to a highly insignificant size. Taken together, this decision-wise analysis raises the possibility that simple adjustments in experimental parameters of the PGG to structurally resemble the specific choice context can make an important contribution towards generalizability.

Table 3: Decision-wise correlations between Task I and Task II

(0) Decision	(1) Group Size	(2) Symmetic	(3) MPCR	(4) Correlation Pooled Sample
a	Large	Yes	0.1	0.0985
b	Large	No	0.1	0.1822**
c	Large	No	0.15	0.2003**
d	Large	Yes	0.2	0.0737
e	Small	No	0.33	0.1713**
f	Small	Yes	0.4	0.1404
g	Large	No	0.42	0.0446
h	Small	No	0.46	0.0956
i	Small	No	0.53	0.0042
j	Small	Yes	0.8	0.0491

Notes: Decision f constitutes the benchmark case.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To further evaluate the potential for generalizability, we now turn to the size of the significant correlation coefficients in table 3. Interpreting their strength requires some point of reference. We propose two reference categories: Correlations between PGG contributions and other abstract tasks that elicit social preferences and pairwise correlations across Task II decisions. The first is a plausible upper limit for the size of correlations between Task I and Task II contributions since behavior in structurally similar games (e.g., a public goods game and a prisoner's

[15] These findings continue to hold, when we adjust p-values to address concerns regarding multiple testing. We employ the method of Dubey, which accounts for the fact that behavior in Task II is highly correlated across decisions. A detailed description of this method can be found in Sankoh et al. (1997)

dilemma) should be more highly correlated than that across structurally less similar decisions. Based on the results of the literature reviewed in Section 2, we find that the degree of generalizability from Task II to Task I is not smaller than that of PGG contributions to behavior in a number of other context-free social preference tasks. The significant correlations in table 3 squarely fall into the range $[r = 0.07; r = 0.41]$ reported in Blanco et al. (2011) and Peysakhovich et al. (2014).[16]

The second reference category, pairwise correlations across single decisions of Task II, relies on data generated by our own experiment and is a more restrictive measure. With the general task structure constant within that task, all variance in individual behavior across single decisions should only reflect changes in experimental parameters. Comparing correlations, we find that the relationship between Task II and Task I is much weaker than that between decisions under changing contribution incentives within Task II. Overall, subjects behave highly consistently across all ten PGG decisions (Cronbach's $\alpha = 0.94$) and correlations between single pairs of decisions range from $r = 0.43$ to $r = 0.85$.[17] Even when contribution incentives strongly differ as, e.g., between decisions b and j, the respective correlation coefficient is larger than any correlation shown in table 3. This apparent difference in size is further corroborated by formal statistical testing: a test for correlated correlation coefficients, as described in Steiger (1980) and Meng et al. (1992), shows that even the highest observed correlation between Task I and Task II (Decision c) is significantly smaller than any correlation observed across different decisions of Task II.

There are at least two potential explanations for the moderate size of correlations in table 3. One is that even the MPCRs in decisions a-d are not sufficiently low to reflect the actual incentives underlying voluntary climate change mitigation efforts in Task I. If so, participants would see Task I and Task II as generally equivalent and the differences in individual behavior between tasks would solely reflect differences in the experimental parameters. In light of the high behavioral consistency throughout Task II, despite substantial parameters changes, such reasoning can only provide a partial explanation of the moderate correlations between tasks. Another potential explanation is that context-specific factors influence individual behavior beyond a generic preference for cooperation. This reasoning is supported by the observation that even when the same participant faces very similar contribution conditions (i.e., sharing money with fellow students in a PGG and a sequential prisoners dilemma), there is only limited evidence for identical behavior at the individual level (Blanco et al., 2011).

Result 3: ***Extensive-margin behavior generalizes better than average behavior. A variation of experimental parameters has little impact on the correlation between free-riding in Tasks I and II.***

So far, we have analyzed behavioral consistency based on comparisons between the (average) amounts contributed to the respective public goods. There is reason to believe, however, that extensive-margin decisions (whether or not to contribute at all) could be determined by different

[16] The fact, that even for these more comparable contribution tasks some correlations are weak to negligible mirrors findings from social psychology (Ross and Nisbett, 2011) which underline that individual behavior is often strongly influenced by situational factors and only to a limited degree attributable to stable traits.

[17] A full correlation table can be found in the Appendix table 6.

factors than the subsequent decision about the size of the contribution (Bergstrom et al., 1986; Smith et al., 1995; Kotchen and Moore, 2007). If so, the previous analysis could have overlooked an aspect of Task II that indeed generalizes to Task I. We therefore repeat the main steps of the previous analysis, now examining extensive-margin behavior.

A first, rough summary measure of the extensive margin is the percentage of decisions in which subjects contribute zero tokens in Task II. Based on this measure, 12.6 percent of subjects are categorized as strict free-riders because they never contribute to the public account. By comparison, 39.3 percent of subjects do not contribute to the public good of climate change mitigation in Task I. While these mean rates of free-riding differ substantially, we now find evidence for similar behavior at the individual level: Free-riding in the two tasks is correlated in a weakly significant way ($r_s = 0.1521$; $p = 0.0783$) when looking at all Task II decisions. There, 59 percent of strict free-riders also do not contribute in the mitigation task. The evidence becomes stronger when we look at distinct decisions within Task II. For the benchmark case, we find a significant correlation ($r_s = 0.1992$; $p < 0.05$) between individual free-riding behavior in decision f and in the mitigation task. For eight out of ten decisions there is a significant ($p < 0.05$) positive correlation in the narrow range from $r_s = 0.1905$ to $r_s = 0.2573$. The smallest insignificant correlation $r_s = 0.1153$ is again found in decision j which is characterized by the highest MPCR.[18]

4.3 The role of subject pool

A considerable number of studies have examined whether conducting experiments with a convenience sample of students affects the conclusions that can be drawn from economic experiments on social preferences (Gächter et al., 2004; List, 2004; Carpenter et al., 2008; Thöni et al., 2012; Anderson et al., 2013; Falk et al., 2013; Belot et al., 2015). The main concern is that students share only a limited range of individual attributes with the general population and, hence, could lack an important determinant of population behavior. It is subject to an ongoing discussion whether this concern mainly applies to level effects (e.g., in our case the size of contributions) or also to treatment effects (Harrison and List, 2004). Figure 1 clearly shows that the average student contributes significantly less in both tasks than the average non-student. Thus, our results conform to prior evidence that the behavior of students can be seen as a lower bound for the extent of pro-sociality one can expect among a more heterogeneous population. But does this significant level effect also imply that more could be learned about voluntary mitigation decisions from conducting a conventional PGG experiment with participants from a more diverse, and therefore more policy relevant, study population? This would only be the case if behavior from PGGs transferred equally well to the mitigation context for students and non-students. The mixed results of the studies reviewed in Section 2 raise the possibility that this is not necessarily the case. For instance, some of the studies - especially those drawing on student subjects (Laury and Taylor, 2008; Benz and Meier, 2008) - have found significant correlations while studies conducted among a more diverse population (Voors et al., 2012) have not detected a significant relationship. Yet, as each of these studies observes contributions to a specific real

[18] A full table containing decision-wise correlations for free-riding can be found in the Appendix table 7.

public good, it is unclear whether their opposing results indeed arise from systematic differences between their respective subject pools. By contrast, we observe participants drawn from two distinct subject pools interacting with the same public good. Hence, we can analyze if correlations differ between those two subject pools.

Result 4: ***For student subjects, behavior in the PGG is more strongly correlated with behavior in the real giving task than for non-student subjects.***

When breaking down our prior analysis by student status, we find that the results reported above are mainly driven by the consistent choices of students. The correlation between average contributions in the PGG and contributions in Task I is slightly larger, yet still insignificant, for students ($r = 0.1531$; $p = 0.3288$). For non-students this correlation is negligible ($r = 0.0312$; $p = 0.7196$). As shown in table 4, this disparity is not driven by a single PGG decision. Instead, irrespective of the parametrization, for non-students all correlations are very low.

Table 4: Decision-wise correlations between Task I and Task II

(0) Decision	(1) Group Size	(2) Symmetric	(3) MPCR	(4) Correlation Non-Students	(5) Correlation Students
a	Large	Yes	0.1	0.0027	0.1689
b	Large	No	0.1	0.1081	0.3723**
c	Large	No	0.15	0.1319	0.3516**
d	Large	Yes	0.2	-0.0184	0.2939*
e	Small	No	0.33	0.0906	0.2964*
f	Small	Yes	0.4	0.0827	0.1455
g	Large	No	0.42	-0.0074	0.0570
h	Small	No	0.46	0.0242	0.1880
i	Small	No	0.53	-0.0452	0.1308
j	Small	Yes	0.8	-0.0719	0.1376

Notes: Decision f is the benchmark case. For student subjects we exclude one apparent outlier shown in figure 1. Including this outlier reduces correlation in size.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

For students, however, there are significant correlations for some of the decisions in Task II. The choice of experimental parameters again influences the strength of these correlations. Only when the MPCR is smaller or the group size is larger than in the benchmark case of decision f, correlations are sizable. This difference between subject pools is robust to accounting for the higher demographic heterogeneity among non-student subjects. By calculating partial correlation coefficients, which hold constant the set of observed characteristics contained in table 2, we still find significant correlations only for student subjects.[19]

An additional analysis of free-riding behavior mirrors these findings. Only students display a (borderline) significant correlation when averaging over all ten decisions ($r_s = 0.2967; p = 0.0534$) of the PGG. Students who free-ride in Task I, on average contribute a significantly smaller fraction of their endowment in Task II (13.35 percent vs. 27.05 percent; Mann-Whitney

[19] Alternative robustness checks yield equivalent results. In a SURE framework, using the same demographic controls, Breusch-Pagan tests reject the hypothesis that residuals are independent for three out of four decisions shown to be significantly correlated in table 4 for student subjects. For non-students this hypothesis cannot be rejected for any decision.

Rank-Sum Test: $p = 0.01$). These results do not carry over to non-students. For them, the correlation between average free-riding in the abstract task and contributing zero in the real contribution task is negligible ($r_s = 0.0511$; $p = 0.6287$). Similarly, free-riding in the real contribution task is unrelated to average contributions in Task II. A decision-wise analysis of free-riding retains the previous result that the correlation structure is largely unaffected by the choice of parameters. For students there is a significant correlation for almost every decision ($r_s = 0.28$ to $r_s = 0.39$), while non-students reveal no significant correlation for any single decision.[20]

4.4 The joint role of task format and individual characteristics

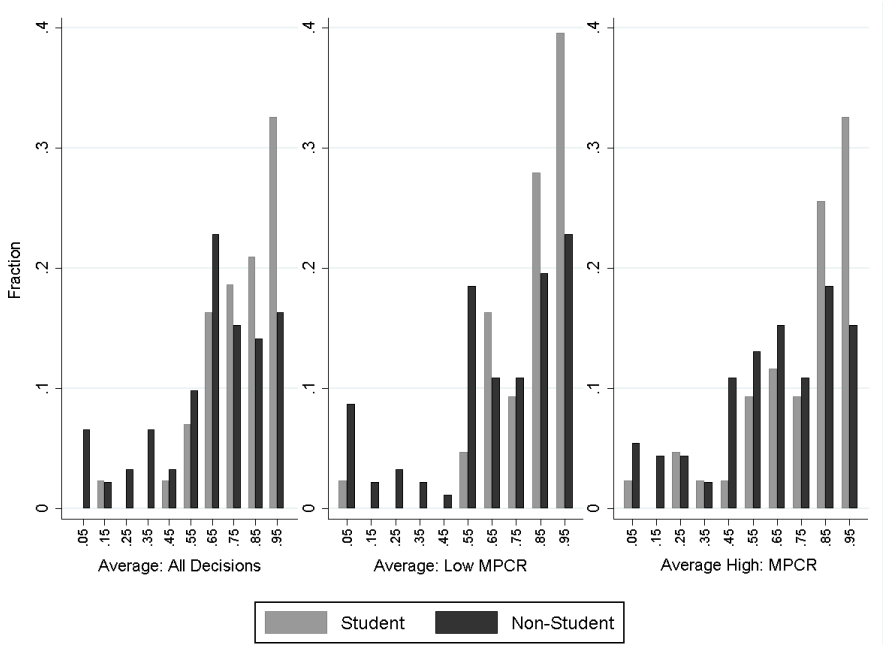
The sections above have highlighted how both the experimental parameters in the PGG and the choice of the subject pool can influence the degree to which results on contribution behavior are readily transferable to the context of voluntary climate change mitigation. In this section we expand these previous results along two dimensions. First, we explore the joint role of subject-pool effects and task format. Second, we look at key attributes beyond student status that could account for subject pool effects. This second step might help to identify specific segments of the population for which PGG behavior is particularly generalizable. If possible, this characterization could provide some guidance when targeting specific study populations, for which one can expect results to be meaningfully interpretable in the mitigation context.

Result 5: Quantitatively, subject pool effects outweigh the effect of game parameters in explaining individual consistency. These differences cannot be attributed to observable characteristics.

As a first step, we define a measure of individual behavioral consistency. By our stylized definition, a pair of choices would count as perfectly consistent if a decision-maker selected identical actions in an identical setting. As a simple measure that conforms with this definition, we calculate the absolute difference between the fractions of endowment contributed in Task I and Task II and subtract it from one. Clearly, whether or not a given decision maker indeed perceives choices in Task I and Task II as equivalent could depend on context specific factors (e.g., game parameters and framing), individual characteristics determining his preferences in each task, and the interaction of these factors (Furr and Funder, 2004). Applied to our experiment, if behavior in both tasks was driven by exactly the same set of individual characteristics and contextual factors did not matter, our measure would be one for the same individual. In contrast, if for the two tasks these factors worked in opposite directions, the measure would tend towards zero.
Figure 3 displays the distribution of this consistency measure for the two distinct subject pools. From left to right, we show three different averages: One average across all ten decisions of Task I, another only for low MPCR (< 0.4) decisions, and the third only for high MPCR (≥ 0.4) decisions.[21] The figure reveals similar patterns as the previous sections, but also highlights the extent of individual heterogeneity. A considerable share of participants conform to our definition

[20] A full table containing decision-wise correlations for free-riding can be found in the Appendix table 7.
[21] Each of these average measures is calculated according the following formula using the notation introduced

Figure 3: Distribution of average consistency



Notes: Histogram displaying the distribution of different average consistency measures by subject pool.

of "perfect consistency". Across the three panels, between 15 and 40 percent of subjects select almost identical contributions in both tasks. Comparing the middle panel to those to its left and right shows that identical choices are most common among students taking the low MPCR decisions. Consistent free-riding accounts for more than half of this fraction. However, especially among non-students, there is also a large group of subjects who reach only a low to medium level of consistency.

In a more refined analysis, we now check whether this heterogeneity can be linked to the variation of individual attributes and contextual factors. The resulting regression model makes use of the full panel structure of our data. For each individual we observe ten decision-wise consistency measures, which is our dependent variable $(1 - |\frac{x_{it}}{\omega} - g_i|)$. Across all 1,350 observed realizations of this variable, we find 118 instances of perfect inconsistency and 335 instances of perfect consistency. The largest part (63.5 percent) of consistent decisions are by subjects who free-ride in both tasks, followed by subjects who contribute half of their endowment (23.9 percent) and full contributors (5.3 percent). This conforms with the findings of others, stating that free-riding is the most stable individual behavior within the same task, across different cooperation tasks and across time (Brosig et al., 2007; Ubeda, 2014). To quantify to what degree behavioral differences in the two tasks are driven by parameter choices and to what degree they are linked to individual characteristics, we estimate different specifications of a random effects tobit model

in Section 3, with g_i denoting the fraction of endowment contributed by individual i in Task I:

$$c_i^z = 1 - \frac{1}{T}\sum_{t}^{T} |\frac{x_{it}}{\omega} - g_i| \tag{2}$$

shown in table 5.

Table 5: Differences in behavior, Task Format and Individual Characteristics

	(1) Consistency	(2) Consistency	(3) Consistency
MPCR	-0.218****	-0.310****	-0.219****
	(-6.17)	(-4.91)	(-6.16)
Non-Student (1=Yes)	-0.233***	-0.282****	-0.242**
	(-3.07)	(-3.49)	(-2.31)
Non-Student*MPCR		0.134*	
		(1.77)	
Age (Years)			0.003
			(0.93)
Male (1=Yes)			-0.101
			(-1.27)
Assets (1=Yes)			0.035
			(0.34)
Years of Education			0.011
			(0.86)
Household Size			-0.019
			(-0.69)
Parent (1=Yes)			-0.230**
			(-2.07)
Stated Risk Aversion (1-11)			-0.004
			(-0.32)
Fear Climate Change (1-7)			-0.009
			(-0.44)
Constant	0.982****	1.016****	0.915****
	(15.26)	(15.10)	(3.70)
Observations	1350	1350	1320
Individuals	135	135	132
Chi^2	47.23	50.35	56.06

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$

Notes: Random effects tobit maximum likelihood estimation to account for censoring from below (0) and above (1). z statistics in parentheses. For each specification the dependent variable is one minus the absolute difference between behavior in Task I and Task II in percentage terms.

In the first specification we jointly estimate the effect of an exogenous variation of the MPCR and moving from a student to a non-student sample. Increasing the MPCR inflates contribution differences between Task I and Task II significantly. Furthermore, for a given MPCR, students display more behavioral consistency than non-students. Quantitatively, the increase in consistency caused by reducing the MPCR from the highest (0.8) to the lowest (0.1) parameterization amounts to approximately two thirds of the effect observed when switching from a non-student to a student subject pool. In specification 2 we show that changes in the MPCR affect students and non-students differently. The weakly significant interaction term indicates that a *ceteris paribus* reduction of the MPCR increases the consistency of students more strongly than

that of non-students. In other words, students react more strongly to changes in contextual factors. In practice, this would mean that a PGG would have to be adapted more strongly when administered to non-students compared to students in order to achieve a similar effect on generalizability. Using only the student status to differentiate between the two subject pools masks a number of individual characteristics that could drive behavioral differences in the two tasks. Thus, specification 3 contains additional controls for individual characteristics. Some of these characteristics, such as gender (Croson and Gneezy, 2009) or age (List, 2004) have been included because they have been shown to influence contribution behavior in standard PGG. Other characteristics such as risk preferences, parenthood, or the fear of climate change could be especially relevant for the decision to contribute to climate change mitigation (Löschel et al., 2013; Diederich and Goeschl, 2014). Thus, these two groups of variables are plausible correlates of context specific preferences in either Task II or Task I. However, with the exception of being a parent, the included characteristics provide no additional information for individual consistency. As the student dummy remains significant and nearly unchanged in size, despite the further control variables, there are likely unobserved individual characteristics that underlie subject-pool differences. Overall, the regression results point out that moving to a more diverse subject pool but retaining the standard task format of a PGG does not necessarily increase the generalizability of results in our context. Subject-pool specific differences have a larger impact on the overall consistency than differences in the parameterization for the range of values we observe.

5 Discussion and conclusion

In the past decades, experiments have started to play an increasingly important role in economic research. In line with this development, there is also a growing interest in drawing on experimental methods and evidence to illuminate concrete policy debates, such as those surrounding climate change mitigation (Bohm, 2003). We agree that in this regard much could be learned from experiments, as they offer a cheap and feasible way to gain insights into the behavioral responses to novel policies within a controlled environment. But for experiments motivated by specific policy issues, generalizability becomes a central issue (Schram, 2005; Sturm and Weimann, 2006).

Our analysis highlights that heterogeneity in mitigation decisions is indeed partly attributable to generic cooperative preferences, but also depends on idiosyncratic factors. Of course, for policy advice the main advantage of experiments lies in their ability to isolate the effects of a particular treatment variation on behavior. Given that in our experiment a considerable fraction of individual mitigation decisions are driven by latent variables not observed in the PGG (especially when using standard parameters), it is not obvious whether treatment effects would be highly transferable between these two settings, in a quantitative and maybe even in a qualitative sense.[22] Clearly, this does not mean that such concerns materialize necessarily for

[22] As highlighted by Kessler and Vesterlund (2015), a discussion about qualitative transferability might be more fruitful. However, even for qualitative treatment effects with an unknown underlying causal mechanism (Heckman and Smith, 1995; Imai et al., 2011) the potential for transferability is hard to assess, because it is not clear which latent factors (common or idiosyncratic) link the treatment variable to the outcome.

all treatment effects of interest. For instance, the qualitative predictions regarding the effects of providing social information have been largely unaffected by the setting under which they were obtained, be it for contributions in abstract laboratory PGG tasks (Bardsley, 2000), in different field settings (Alpizar et al., 2008; Shang and Croson, 2009), or in the specific context of mitigation decisions (Allcott, 2011).

Importantly, we do not see our finding as discarding the application of (abstract or context-specific) experiments to questions of climate change policy. Rather to the contrary, they call for more experimentation, in the spirit of the arguments raised in Falk and Heckman (2009). Only by obtaining further experimental evidence one can shed additional light on the conditions under which one can safely assume a high level of generalizability. We make a first step into this direction within our own framework and explore two potential shifters of generalizability. Our first treatment variation suggests that the link between PGG behavior and mitigation decisions can be strengthened by bringing the experimental parameters closer to the context of interest. For PGG with a low MPCR the correlation between behavior in both tasks increases, sometimes even substantially. Consequently, in the limit, the best laboratory equivalent to individual mitigation behavior might well turn out to be the standard dictator game in which the dictator's private return of contributing is zero. So far, there is only limited experimental evidence on contribution behavior from PGG under conditions of very low MPCR (Weimann et al., 2012). While some general patterns persist, there is also some emerging evidence that well known mechanisms for fostering cooperation such as peer-punishment (Xu et al., 2013) are much less effective given a reduced MPCR. Further research in this direction could be of great interest for those who wish to study the behavioral mechanisms of cooperation in the context of climate change.

From our second treatment variation we derive more ambivalent conclusions, regarding questions of generalizability. If it was a central aim to make statements about the level of cooperation, the use of a convenience student sample could be somewhat misleading. We replicate earlier findings that student behavior is only a lower bound for the cooperative behavior that can be expected in a population with broader demographic heterogeneity. On the other hand, we show that students are more responsive to changes in experimental parameters (or conversely less responsive to differences between the tasks) and consequently display a higher consistency between the different decision tasks. Thus, sampling from the general population, with the aim to draw from a more representative subject pool might impose stronger demands on the experimental design. The higher diversity of the subject pool might not only call for a larger sample size but also for additional treatment variations.

Clearly, our experiment is only a first step towards understanding generalizability in the narrow context of climate change mitigation. The larger question, namely whether social preference tasks are generally external valid, cannot be resolved on its basis as our results are, by design, context-dependent. A relevant extension of our design would replace Task I with an actual measurement of voluntary mitigation behavior in a field environment. Such a measure would differ from Task I along several dimensions. Mitigation decisions outside the lab context require individuals to use their own money instead of an experimental endowment, are not scrutinized by an experimenter but instead (in some cases) by the social environment and are often bundled

with other attributes of a consumption decision. Each of these shift parameters reduces the artificiality of Task I relative to Task II. It is left for further research to assess how this would affect conclusions about generalizability.

Acknowledgement: *We thank the panel SÖF (Sozial Ökologische Forschung) within the BMBF Germany (Bundesministerium für Bildung und Forschung) who provided funds for conducting this study. We are grateful to our student research assistants Dennis Daseking, Florian Bisinger, Korbinian Dress, Elisabeth Dorfmeister and Johanna Köhling who helped with sampling and conducting this study.*

References

O. Al-Ubaydli and J. A. List. On the generalizability of experimental results in economics. In G. R. Fréchette and A. Schotter, editors, *Handbook of Experimental Economic Methodology*, pages 420–463. Oxford University Press, 2015.

H. Allcott. Social norms and energy conservation. *Journal of Public Economics*, 95(9):1082–1095, 2011.

F. Alpizar, F. Carlsson, and O. Johansson-Stenman. Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics*, 92(56):1047–1060, 2008.

J. Anderson, S. V. Burks, J. Carpenter, L. Götte, K. Maurer, D. Nosenzo, R. Potter, K. Rocha, and A. Rustichini. Self-selection and variations in the laboratory measurement of other-regarding preferences across subject pools: evidence from one college student and two adult samples. *Experimental Economics*, 16(2):170–189, 2013.

N. Bardsley. Control without deception: Individual behaviour in free-riding experiments revisited. *Experimental Economics*, 3(3):215–240, 2000.

M. Belot, R. Duch, and L. Miller. A comprehensive comparison of students and non-students in classic experimental games. *Journal of Economic Behavior & Organization*, 113:26–33, 2015.

M. Benz and S. Meier. Do people behave in experiments as in the field? evidence from donations. *Experimental Economics*, 11(3):268–281, 2008.

T. Bergstrom, L. Blume, and H. Varian. On the private provision of public goods. *Journal of Public Economics*, 29(1):25–49, 1986.

M. Blanco, D. Engelmann, and H. T. Normann. A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2):321–338, 2011.

P. Bohm. Experimental evaluations of policy instruments. In Karl-Göran Mäler and Jeffrey R. Vincent, editor, *Handbook of Environmental Economics*, pages 437–460. Elsevier, 2003.

K. A. Brekke and O. Johansson-Stenman. The behavioural economics of climate change. *Oxford Review of Economic Policy*, 24(2):280–297, 2008.

K. Brick and M. Visser. What is fair? An experimental guide to climate negotiations. *European Economic Review*, 74(0):79–95, 2015.

J. Brosig, T. Riechmann, and J. Weimann. Selfish in the end?: An investigation of consistency and stability of individual behavior. *FEMM Working Paper No. 05*, 2007.

C. Camerer. The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. In G. R. Fréchette and A. Schotter, editors, *Handbook of Experimental Economic Methodology*, pages 249–296. Oxford University Press, 2015.

F. Carlsson and O. Johansson-Stenman. Behavioral economics and environmental policy. *Annu. Rev. Resour. Econ.*, 4(1):75–99, 2012.

J. Carpenter, C. Connolly, and C. K. Myers. Altruistic behavior in a representative dictator experiment. *Experimental Economics*, 11(3):282–298, 2008.

A. Chaudhuri. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, 14(1):47–83, 2011.

R. Croson and U. Gneezy. Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474, 2009.

A. C. de Oliveira, R. T. Croson, and C. Eckel. The giving type: Identifying donors. *Journal of Public Economics*, 95(56):428–435, 2011.

J. Diederich and T. Goeschl. Willingness to pay for voluntary climate action and its determinants: Field-experimental evidence. *Environmental and Resource Economics*, 57(3):405–429, 2014.

C. C. Eckel and P. J. Grossman. Altruism in anonymous dictator games. *Games and economic behavior*, 16(2):181–191, 1996.

F. Exadaktylos, A. M. Espín, and P. Brañas Garza. Experimental subjects are not different. *Nature: Scientific reports*, 3, 2013.

A. Falk and J. J. Heckman. Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952):535–538, 2009.

A. Falk, S. Meier, and C. Zehnder. Do lab experiments misrepresent social preferences? The case of self-selected student samples. *Journal of the European Economic Association*, 11(4): 839–852, 2013.

E. Fehr and A. Leibbrandt. A field study on cooperativeness and impatience in the tragedy of the commons. *Journal of Public Economics*, 95(9-10):1144–1155, 2011.

P. J. Ferraro and C. A. Vossler. The source and significance of confusion in public goods experiments. *The BE Journal of Economic Analysis & Policy*, 10(1):1935–1682.2006, 2010.

U. Fischbacher. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178, 2007.

R. Furr and D. C. Funder. Situational similarity and behavioral consistency: Subjective, objective, variable-centered, and person-centered approaches. *Journal of Research in Personality*, 38(5):421–447, 2004.

S. Gächter, B. Herrmann, and C. Thöni. Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence. *Journal of Economic Behavior & Organization*, 55(4):505–531, 2004.

M. M. Galizzi and D. Navarro Martinez. On the external validity of social-preference games: A systematic lab-field study. *Workin Paper Series*, 2015.

J. K. Goeree, C. A. Holt, and S. K. Laury. Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. *Journal of Public Economics*, 83(2):255–276, 2002.

J. M. Gowdy. Behavioral economics and climate change policy. *Journal of Economic Behavior & Organization*, 68(3):632–644, 2008.

B. Greiner. An online recruitment system for economic experiments. *Journal of the Economic Science Association*, (1), 2015.

E. Gsottbauer and J. van den Bergh. Environmental Policy Theory Given Bounded Rationality and Other-regarding Preferences. *Environmental and Resource Economics*, 49(2):263–304, 2011.

G. W. Harrison and J. A. List. Field experiments. *Journal of Economic Literature*, 42(4): 1009–1055, 2004.

J. J. Heckman and J. A. Smith. Assessing the case for social experiments. *The Journal of Economic Perspectives*, 9(2):85–110, 1995.

D. Houser and R. Kurzban. Revisiting kindness and confusion in public goods experiments. *The American Economic Review*, 92(4):1062–1069, 2002.

K. Imai, L. Keele, D. Tingley, and T. Yamamoto. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(04):765–789, 2011.

R. M. Isaac and J. M. Walker. Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics*, 103(1):179, 1988.

J. Kessler and L. Vesterlund. The external validity of laboratory experiments: The missleading emphasis on quantitative effects. In G. R. Fréchette and A. Schotter, editors, *Handbook of Experimental Economic Methodology*, pages 391–407. Oxford University Press, 2015.

M. J. Kotchen and M. R. Moore. Private provision of environmental public goods: Household participation in green-electricity programs. *Journal of Environmental Economics and Management*, 53(1):1–16, 2007.

S. K. Laury and L. O. Taylor. Altruism spillovers: Are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good? *Journal of Economic Behavior & Organization*, 65(1):9–29, 2008.

J. O. Ledyard. Public Goods: A Survey of Experimental Research. In J. Kagel and A. Roth, editors, *Handbook of experimental economics*. Princeton University Press, Princeton, 1995.

S. D. Levitt and J. A. List. What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives*, 21(2):153–174, 2007.

S. D. Levitt and J. A. List. Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53(1):1–18, 2009.

J. A. List. Young, Selfish and Male: Field evidence of social preferences. *The Economic Journal*, 114(492):121–149, 2004.

A. Löschel, B. Sturm, and C. Vogt. The demand for climate protection Empirical evidence from Germany. *Economics Letters*, 118(3):415–418, 2013.

X.-L. Meng, R. Rosenthal, and D. B. Rubin. Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1):172, 1992.

M. Milinski, D. Semmann, H.-J. Krambeck, and J. Marotzke. Stabilizing the earths climate is not a losing game: Supporting evidence from public goods experiments. *PNAS*, 103(11): 3994–3998, 2006.

M. Milinski, R. D. Sommerfeld, H.-J. Krambeck, F. A. Reed, and J. Marotzke. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *PNAS*, 105 (7):2291–2294, 2008.

W. D. Nordhaus. To slow or not to slow: The economics of the greenhouse effect. *The Economic Journal*, pages 920–937, 1991.

D. Nosenzo, S. Quercia, and M. Sefton. Cooperation in small groups: The effect of group size. *Experimental Economics*, 18(1):4–14, 2015.

A. Peysakhovich, M. A. Nowak, and D. G. Rand. Humans display a cooperative phenotype that is domain general and temporally stable. *Nat Commun*, 5, 2014.

L. Ross and R. E. Nisbett. *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers, 2011.

A. J. Sankoh, M. F. Huque, and S. D. Dubey. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*, 16(22):2529–2542, 1997.

A. Schram. Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, 12(2):225–237, 2005.

J. Shang and R. Croson. A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal*, 119(540): 1422–1439, 2009.

J. F. Shogren and L. O. Taylor. On behavioral-environmental economics. *Review of Environmental Economics and Policy*, 2(1):26–44, 2008.

V. H. Smith, M. R. Kehoe, and M. E. Cremer. The private provision of public goods: Altruism and voluntary giving. *Journal of Public Economics*, 58(1):107–126, 1995.

C. Starmer and R. Sugden. Does the random-lottery incentive system elicit true preferences? An experimental investigation. *American Economic Review*, 81(4):971–78, 1991.

J. H. Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87 (2):245, 1980.

B. Sturm and J. Weimann. Experiments in environmental economics and some close relatives. *Journal of Economic Surveys*, 20(3):419–457, 2006.

A. Tavoni, A. Dannenberg, G. Kallis, and A. Löschel. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *PNAS*, 108(29):11825–11829, 2011.

C. Thöni, J.-R. Tyran, and E. Wengström. Microfoundations of social capital. *Journal of Public Economics*, 96(78):635–643, 2012.

P. Ubeda. The consistency of fairness rules: An experimental study. *Journal of Economic Psychology*, 41(0):88–100, 2014.

L. Venkatachalam. Behavioral economics for environmental policy. *Ecological Economics*, 67 (4):640–645, 2008.

L. Vesterlund. Voluntary giving to public goods: Moving beyond the linear VCM. In J. Kagel and A. Roth, editors, *Handbook of experimental economics*. Priceton University Press, Priceton, 2014.

M. Voors, T. Turley, A. Kontoleon, E. Bulte, and J. A. List. Exploring whether behavior in context-free experiments is predictive of behavior in the field: Evidence from lab and field experiments in rural Sierra Leone. *Economics Letters*, 114(3):308–311, 2012.

J. Weimann, J. Brosig, H. Hennig-Schmidt, C. Keser, and C. Stahr. Public-good experiments with large groups. *Magdeburg University Working Paper 9/2012*, 2012.

B. Xu, C. B. Cadsby, L. Fan, and F. Song. Group size, coordination, and the effectiveness of punishment in the voluntary contributions mechanism: An experimental investigation. *Games*, 4(1):89–105, 2013.

J. Zelmer. Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6(3): 299–310, 2003.

6 Appendix: Supplementary tables and regressions

6.1 Correlation table task II: Decision a.-j.

Table 6 contains the correlation coefficients for each pair of decisions made in task Task II.

Table 6: Correlation Matrix of Task II decisions

Decisions	a	b	c	d	e	f	g	h	i	j
a	1.000									
b	0.681	1.000								
c	0.697	0.849	1.000							
d	0.706	0.731	0.696	1.000						
e	0.674	0.716	0.701	0.642	1.000					
f	0.617	0.691	0.598	0.696	0.758	1.000				
g	0.658	0.626	0.572	0.749	0.516	0.611	1.000			
h	0.587	0.564	0.480	0.613	0.597	0.655	0.691	1.000		
i	0.555	0.494	0.528	0.583	0.579	0.559	0.613	0.721	1.000	
j	0.467	0.431	0.436	0.544	0.469	0.504	0.588	0.762	0.625	1.000

6.2 Correlations free-riding

Table 7 contains Spearman correlation coefficients between free-riding in Task I and Task II. For the pooled sample (4) there are significant correlations for eight out of ten Task II decisions. These mainly reflect consistent free-riding among student subjects (6).

Table 7: Spearman correlations between free-riding in the real and in the abstract context for all 10 decisions

(0)	(1)	(2)	(3)	(4)	(5)	(6)
Decision	Group Size	Symmetry	MPCR	Correlation	Correlation Non-Students	Correlation Students
a	Large	Sym	0.1	0.2085**	0.1196	0.3486**
b	Large	Asym	0.1	0.1924**	0.0919	0.3603**
c	Large	Asym	0.15	0.2221***	0.1196	0.3908***
d	Large	Sym	0.2	0.2573***	0.1738	0.3841**
e	Small	Asym	0.33	0.1261	0.0067	0.3072**
f	Small	Sym	0.4	0.1992**	0.13	0.2969*
g	Large	Asym	0.42	0.2051**	0.1201	0.3341**
h	Small	Asym	0.46	0.1905**	0.0378	0.3841**
i	Small	Asym	0.53	0.2133**	0.11	0.3812**
j	Small	Sym	0.8	0.1153	-0.0045	0.2861*

Notes: Decision f constitutes the benchmark case.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

6.3 Regression results task II

Table 8 displays results from a random effects tobit regression with the fraction of endowment contributed as the dependent variable. The most basic specification (1) corroborates a positive and significant relationship between contributions and the internal return, external return, group size in each decision of Task II. Furthermore, non-students contribute higher amounts. These relationships are robust to controlling for further demographic variables and attitudes in specification (2).

Table 8: Contributions Abstract Public Good Game and Demographic Variables

	(1) Contributions	(2) Contributions
Non-Student (1=Yes)	0.333****	0.225*
	(3.93)	(1.94)
Internal Return	0.029****	0.029****
	(6.98)	(6.88)
External Return	0.012****	0.013****
	(3.89)	(4.04)
Group Size	0.021****	0.021****
	(5.79)	(5.80)
Age (Years)		0.007*
		(1.82)
Male (1=Yes)		-0.031
		(-0.36)
Assets (1=Yes)		-0.245**
		(-2.46)
Years of Education		-0.008
		(-0.57)
Household Size		0.017
		(0.56)
Number of Children		0.049
		(1.00)
Fear Climate Change (1-7)		-0.036
		(-1.51)
Constant	-0.424****	-0.304
	(-4.82)	(-1.12)
Observations	1350	1320
Individuals	1350	1320
Prob > Chi^2	0.000	0.000

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$

Notes: Random effects tobit maximum likelihood estimation to account for censoring from below (0) and above (1). z statistics in parentheses.