Growth and inequality in public good provision — an extended replication

Hauke Roggenkamp[a,b]

[a] *Helmut Schmidt University, Holstenhofweg 85, Hamburg, 22043,*

[b] *University of St. Gallen, Torstrasse 25, St. Gallen, 9000,*

Abstract

You can find the most recent version of this paper here. The abstract follows at later point.

Keywords: Replication study, Non-convenience sample, Open science, Dynamic public good game, Online experiment, Generalizability

1. Introduction

Today's actions are tomorrow's result. There are many settings in which current decisions affect future outcomes and with it, future decision spaces. Opting for environmental friendly policies today not only reduces carbon dioxide omissions immediately but also helps us to reach the Paris climate targets tomorrow. Deferring these policies, may not necessarily prevent us from reaching these targets, but it makes it more difficult in the future. Hence, today's actions (or the omission thereof) not only affect intermediate outcomes but also the number of paths one can choose to reach certain goals.

Standard public good games—although often intended to inform climate policies[1]—miss these temporal interdependencies simply because participants have the same set of actions in each period. Accordingly, participants' actions in a given period do not affect their number of actions in subsequent periods. A game implemented by Gächter, Mengel, Tsakas & Vostroknutov (2017) (hereafter, GMTV) as well as Stefan Große (2011, unpublished) incorporated interdependencies into a *dynamic* public good game. Here, participants' actions in a given period affect their number of actions in subsequent periods. Because there is surprisingly little experimental research on these interdependencies[2], I replicated one of their treatments.

Goeschl et al. (2020) find that standard public good games do not generalize to real-world climate action. Because the dynamic setting has a more realistic property, one wonders whether it is better suited to inform public policy. To answer this question, I ran the experiment with different samples[3] online, that is, with participants who make their decisions in an environment whcih is more natural than the lab. In addition, I observed the participants' behavior in voluntary climate actions (VCA). This yielded a setting similar to Goeschl et al. (2020)'s which allows me to analyze how behavior in the abstract game translates into real-world action across samples.

Arechar et al. (2018) conducted public good games in the lab and on MTurk to draw lessons from online experimentation. This study extends this literature (see also Goodman and Paolacci, 2017) by focusing on a completely inexperienced sample that has not been exposed to interactive experiments before. This

*Corresponding author
Email address: `hauke.roggenkamp@unisg.ch` (Hauke Roggenkamp)

[1] See Goeschl et al. (2020, p. 1) for numerous references.

[2] One exemption are Eichenseer and Moser (2019), who built on GMTV's design do investigate leadership.

[3] GMTV ran their experiments with students in Nottingham, England.

required me to design a robust (and thus, more complex) software to minimize attrition. In exchange, the online setting allowed me to also collect paradata to assess the fluency and feasibility of the experiment. This article reports on both topics: the robust design as well as the feasibility.

Taken together, this study makes three contributions. First, it replicates parts of GMTV's original experiment and highlights the importance of pure replications. Second, it shows that logistically complex online experiments are feasible for samples other than students or clickworkers. Third, it shows that findings from abstract games do not generalize well and even worse with the representative sample. After reporting the methods, this paper is organized along these findings.

2. Methodology

In the terminology of Hamermesh (2007), I ran both a *pure* as well as a *scientific replication*[4] of one treatment of GMTV's dynamic public good game. The pure replication re-analyzes the original data. Appendix A documents errors I identified in the original paper. The scientific replication, where I utilize a different sample drawn from a different population in a different situation, is described in the following sections.

2.1. Experimental Design

As in the NOPUNISH 10 Period treatment of GMTV, I ran sessions with groups of four ($i \in I = \{1, 2, 3, 4\}$), an initial endowment of $N_i^1 = 20$ tokens[5], $T = 10$ periods, a private account with a return of 1 and a group account with a return of 1.5 ($\Rightarrow$ MPCR$\equiv \frac{1.5}{4}$). With i's contribution in period t being c_i^t, the model looks as follows:

$$N_i^{t+1} = N_i^t - c_i^t + \frac{1.5}{4}\sum_{j=1}^{4} c_j^t$$

What makes the game *dynamic*? Instead of receiving fresh endowments every period, participants received one endowment only at the beginning of the first period. A participant's endowment in the second period is the wealth he or she accumulated in the first period. A participant's endowment in the third period is the wealth he or she accumulated in the first two periods. And so on. Hence, a decision in one period has consequences on future endowments and, ultimately, growth paths. For this reason, the game is described as a *dynamic* public good game

2.2. Voluntary Climate Action (VCA)

Like GMTV we employ a real giving task after the abstract game. In contrast to GMTV and like Goeschl et al. (2020), we employed a VCA, where they could donate any amount their earnings to offset carbon dioxide (that is, retirine emission permits from the EU ETS).[6] To ensure that each participant had the same basic level of information about the impact of their decision, I provided some basic information about the mechanism. The information also highlighted that the mitigation came into operation on a European level. Finally, I informed the participants that the documentation of individual and aggregate contributions were to be posted immediately after the experiments online. To avoid privacy or social image concerns, participants learned their unique and random IDs, which they needed to identify their individual contributions. The document certified that their contributions have been used to offset 1.82 tons of carbon dioxide emissions.

[4] Parsons et al. (2022) use the term of a *conceptual replication* which means the same.

[5] A token was worth 0.05 Euros.

[6] Importantly, Goeschl et al. (2020) made the VCA decision with a fresh endowment *before* they played the abstract game. I deviate from their procedure to match GMTV's procedure.

2.3. Sample

I recruited the participants from the so called *HamburgPanel* using HROOT (Bock et al., 2014). The panel is provided by the University of Hamburg's Research Laboratory, which used a randomized last digits approach to build the panel while drawing from the population of citizens of Hamburg, Germany. Because the sample was exhausted at one point, I also recruited students from the University of Hamburg.

At the time I conducted the experiment, the more representative sample was not familiar with interactive experiments. In fact, I ran the first interactive group experiment with this sample. The students, in contrast, were used to this kind of experiments. Recruiting them, I excluded those who have participated in a public goods game before. As a consequence, nonnaiveté is unlikely to affect the validity of the experiment (Goodman and Paolacci, 2017, p. 204).

Throughout this paper, I will compare results of my experiment with the results of GMTV's NOPUNISH 10 Period treatments. I am thus, referring to three different samples utilized at two points in time: the University of Nottingham's students (in late 2012), Hamburg's citizens and the University Hamburg's students (both in July 2021). Table 1 shows how they compare with respect to a few properties.

Table 1: Sample Properties

	Dependent variable:					
	Female	Age	Trust	Meritocracy	Government	Equality
Hamburg Citizens	0.48^{***}	47.58^{***}	3.90^{***}	3.79^{***}	3.44^{***}	4.17^{***}
	(0.07)	(1.21)	(0.19)	(0.21)	(0.19)	(0.23)
Hamburg Students	0.08	-21.41^{***}	-0.58^{**}	0.23	0.71^{***}	0.12
	(0.09)	(1.63)	(0.25)	(0.28)	(0.26)	(0.30)
Nottingham Students	−0.10	-15.85^{***}	−0.03	1.65^{***}	0.87^{***}	−0.35
	(0.09)	(1.52)	(0.23)	(0.26)	(0.24)	(0.28)
Observations	208	208	208	208	208	208
R^2	0.02	0.47	0.04	0.20	0.06	0.02
Residual Std. Error	0.50	8.75	1.35	1.52	1.40	1.63
Note:						*p<0.1; **p<0.05; ***p<0.01

Describe sample properties here.

2.4. Software

The experiment was logistically complex for several reasons. First, the sample was inexperienced. Second, the experiment was interactive and synchronous. Third, the underlying game was dynamic and interdependent. This makes dropouts not only more likely but also more expensive, which is why attrition was a major concern implementing the experiment.

I chose oTree (Chen et al., 2016) to implement the experiment because it is open-source, well documented and very flexible. Its Bootstrap (a powerful frontend toolkit) integration allowed me to make the graphical user interface interactive, appealing and easy to navigate. The Highcharts library made it easy to visualize results and to communicate dynamics. Insofar, oTree served a good tool to enhance the participants' user experience and thus, to make dropouts less likely. Furthermore, the oTree code snippets made it possible to handle dropouts.

Which features were required to handle dropouts? First, participants had to be matched to form a group *after* comprehension questions were answered successfully. Importantly, participants were grouped by the order they answered these questions to reduce waiting times. While waiting for other players to form the group, the participants saw a wait-page informing them that they are waiting for other participants to arrive and that they do not have to wait for longer than 10 minutes. The screen also informed them that they would receive a *patience bonus* of one Euro after the expiration of that time (or what was left of it). Second, participants only had 10 minutes to make the first contribution and 4 minutes for the remaining contributions. After this time expired, participants were replaced by bots that made random contributions. In this case, the remaining group members were informed about the replacement. Both features were implemented to limit wait times and boredom for other participants. Section 3.2 shows that the first feature became effective in some cases, wheres the second feature did not.

2.5. Procedure

Participants entered the experiment at appointed times remotely from home. They first saw a welcome screen. After agreeing to the privacy policy, they could proceed to the instructions individually. Having read these instructions, each participant has also seen a demo-screen explaining the user interface. Before proceeding, they had to answer six comprehension questions correctly. Subsequently, they saw a waiting screen until they could be matched with three other participants, who have answered the comprehension questions correctly. Once matched, they were exposed to the decision screen over ten periods. At the end of the last period, participants saw results of all periods. Subsequently, they made their VCA decision, before I elicited risk preferences (Holt and Laury, 2002) and finished with GMTV's questionnaire.

While I tried to stick to GMTV's protocol as close as possible, I deviated in a few aspects. First, the instructions were German and also covered topics inherent to the online setting (dropouts and bots, for instance). Second, I used another software (oTree instead of zTree). Third, GMTV gave participants the opportunity to donate to *Doctors without Borders* whereas we offered carbon dioxide offsets. Fourth, the graphical user interface looked different.

The experiment lasted around 25 minutes on average. The earnings averaged 11.23 Euros (sd = 4.85).[7]

2.6. Pre-registration

Parts of the analyses were pre-registered in the American Economic Association's RCT Registry (Berlemann et al., 2021). In addition, I pre-registered the exact analyses I planned to run when I designed designed the experiment on GitHub.[8]. GitHub is a website and cloud-based service where developers—and researchers alike—can store and manage their code. The service is based on Git[9] and designed for version control. Importantly, changes are timestamped and can be tracked. Moreover, one can create branches, that is, duplicates of code – either to work collaborative or to archive a certain state. Accordingly, I archived my analyses scripts and created a branch a few days *before* we collected data.

The analysis that makes the most sense now that I have collected the data is a little different though: Because the more representative subject pool was exhausted earlier than expected, I recruited students unwantedly and changed parts of the scripts.[10] However, because the originally planned analysis code is archived and reproducible, these changes are transparent and easy to assess.

3. Results

3.1. Pre-registered GMTV Replication

3.1.1. Contribution Behavior

First, I ask whether the samples differ with respect to their initial contributions to the public good. Is our replication sample more pro-social than the original sample? Figure 1 reveals that it is not. The

[7] This values include earnings from the incentivized risk elicitation task that is not part of the analysis.

[8] https://github.com/Howquez/coopUncertainty/tree/July21Replication/analysis/reports/rmd to run the code, you need to executed the .Rmd files in this repository in the order that is indicated by its file names.

[9] Git is a specific open-source version control system developed in 2005.

[10] The source code of this document contains the analysis and can be found here: https://github.com/Howquez/coopUncertainty/blob/main/analysis/quarto/paper.qmd

distributions of both samples look fairly similar. Both samples contributed 10 tokens, that is, 50% of their endowments on average (median and mean).[11] Moreover, both samples' initial contributions resemble initial contributions participants usually make in the standard game with partner matching.[12] However, in the dynamic game presented here, we are particularly interested in the subsequent periods because differences add up exponentially. Do the two groups remain similar over the course of time?

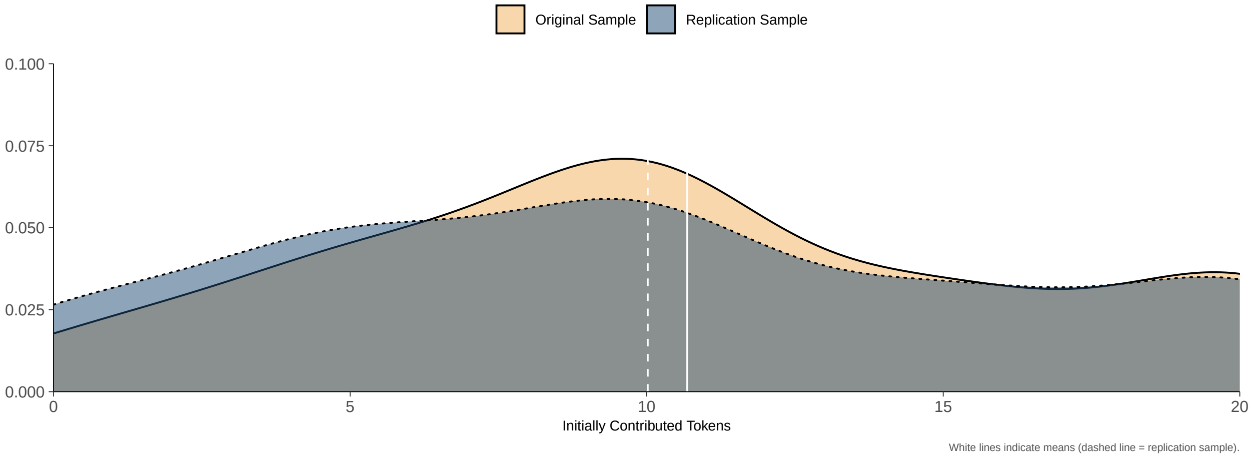

Figure 1: Individual contributions to the dynamic public good in the first period

In particular, do the two samples' contributions follow the same path over the 10 periods they played? The answer is *no.* Figure 2 illustrates that the samples make similar contributions at the beginning and the end of the game but behave differently in between. More precisely, the left panel–depicting the average contributions in absolute terms–shows that the original sample contributed substantially more than the replication sample *in all but the first and last period.* For this reason, the original sample's behavior differs from the replication sample's behavior in two aspects: it contributes more and exhibits a considerable drop in the last period (whereas the replication sample's contributions flatten).

Note that increasing contributions over time imply increasing endowments over time. Hence, absolute contributions do not us much about the willingness to cooperate. For this reason, the right panel in Figure 2 shows the average *share of endowments contributed* over time. Both samples exhibit a similar pattern: their share of endowments contributed declined and did not stabilize. However, both samples also differ with respect to one aspect: the replication sample's share of contributions declines faster.

Again, both samples' behavior resembles the contributions participants usually make in the standard game with partner matching: contributions equal approximately half of endowments in the very first period and decrease to around 10% of endowments by the last period.[13] In the dynamic game presented here, however, different paths lead to different levels of wealth – even if they share the same start- and end-points. I am thus, more interested in the contributions' implications for wealth generation and growth.

3.1.2. Wealth Creation

How do the different contribution-paths translate into wealth?[14] Given that the original sample contributed more in most of the periods, one would expect the respective groups to be considerably more wealthy.

[11] The two-sided rank sum test (comparing differences between samples) yields a p-Value of 0.3926 for the mean contribution in first round of the game.

[12] See Figure 3B in Fehr and Gächter (2000) (p.989), for instance.

[13] The right panel is thus, comparable to the visualizations *and results* in the standard game. See, for instance, Figure 1B in Fehr and Gächter (2000) (p.986).

[14] To measure wealth and growth, I define a variable called *stock* which sums the endowments of all participants in a given group at the end of the round (that is, after the contributions have been made, multiplied and redistributed).

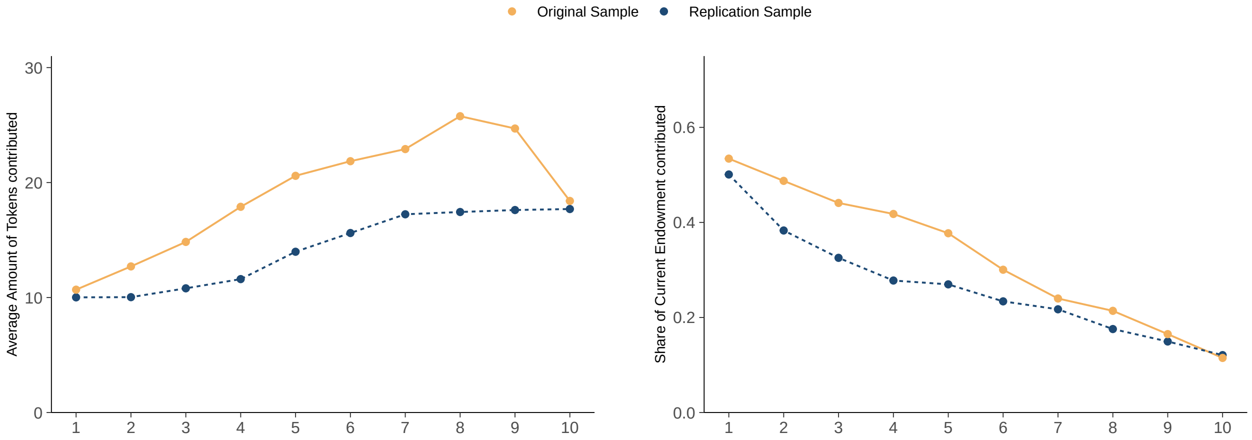

Figure 2: The average amount of tokens contributed over time in treatments.

Figure 3 indicates just that. The grey lines show that an average group in the original sample accumulated about 478 tokens. In contrast, an average group in the replication sample accumulated about 380 tokens. This difference is insignificant at conventional levels[15] though.

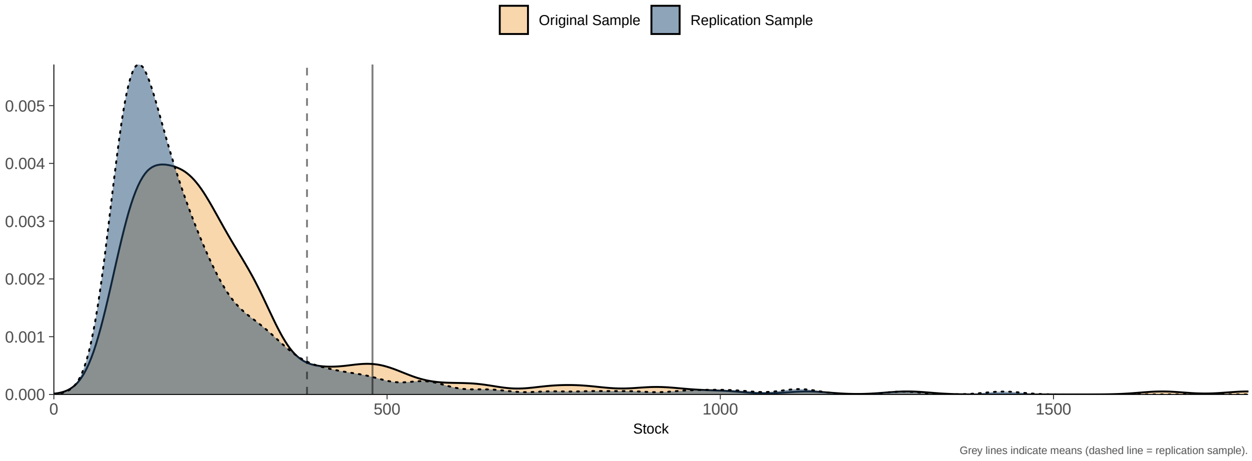

Figure 3: Groups' income at the end of the game

Although there clearly is growth, groups do not realize the maximal potential efficiency: under full cooperation, a group can accumulate at least 4613 tokens or EUR 230. This is depicted in the left panel of Figure 4, where one can see the average wealth over time by sample. The panel illustrates for both samples that growth was continuous and surprisingly linear, given the exponential character of the game's design. To sum up, the contribution behavior differed between samples. In contrast, neither the eventual wealth nor the corresponding growth paths differed. Differences in contribution behavior did, thus, not translate to significantly different wealth outcomes.

Why? Perhaps because the heterogeneity within samples and across groups has been too large to *detect* a significant difference. The right panel of Figure 4 depicts heterogeneity: In the replication sample, the richest group earned 1425 tokens (which is about 1781% of the initial endowment) whereas the poorest group

[15] The two-sided rank sum test (comparing differences between samples) yields a p-Value of 0.1356 for the mean stock in last round of the game.

ends up with 92 tokens (115%). More broadly, the replication sample is characterized by inequality between groups ($SD_{Replication} = 336.06$). The same holds true for the original sample ($SD_{Original} = 393.58$). Hence, the heterogeneity across groups does not differ between samples, which is remarkable because the replication sample was drawn from a more heterogeneous (non-convenience sample). Does it differ within groups?

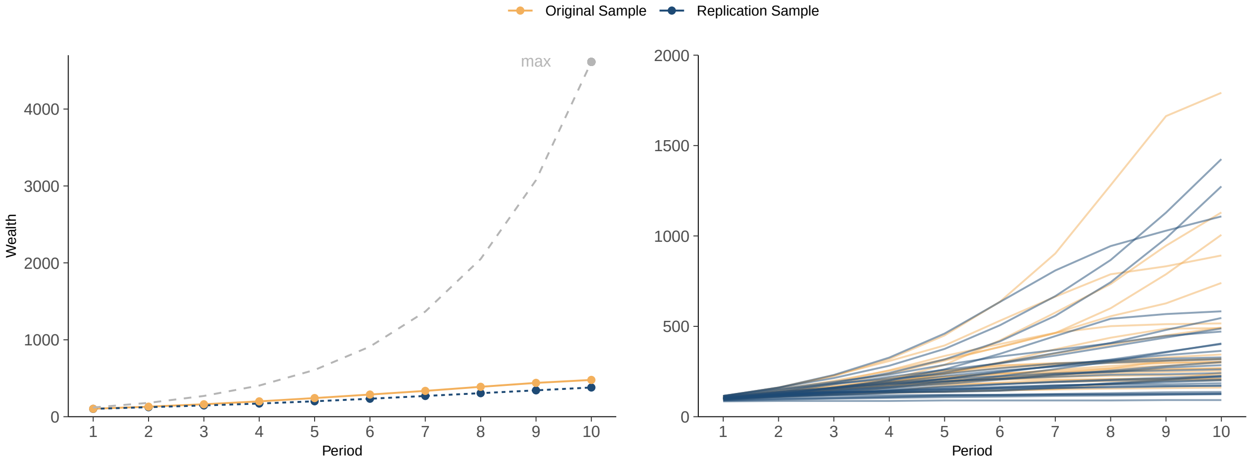

Figure 4: Average wealth over time across samples.

3.1.3. Inequality

Given the different samples and the possibility of endogenous growth–which essentially is the main feature of the game–I ask whether and how the inequality grows *within* groups. Figure 5 illustrates that inequality did grow: at the end of the game, the original and the replication groups exhibit an average Gini coefficient of 0.23 and 0.22, respectively.[16] Because every participant started with the same initial endowment (in *Period 0*, so to speak), every group started equally–with a Gini coefficient equaling zero.

Figure 6 shows that this initial state of equality ended with the first period already: both samples exhibit a stark incline in inequality before the second period started. From then on, the respective Gini coefficients grew slowly but continuously – for both samples.

Result 1. *The NOPUNISH 10 treatment of GMTV can be replicated because the replication data resemble the original data with respect to initial and final contributions, wealth and growth as well as inequality.*

This is remarkable given the different sample and language, the different software and user interface as well as the online setting during the COVID19 pandemic. The result suggests that, by and large, the sum of these factors did not affect people's preferences towards cooperation.

3.2. Online Feasibility

How did the participants, who have never participated in an online group experiment before, cope with the situation? Moreover, did participants understand the unfamiliar setting they found themselves in? While the answer to the former question requires more thought, the answer to the latter simply is *yes*: 67 out of 116 answered with *"yes"* when I asked them. Another 44 answered with *"rather yes"* while nobody indicated that he or she did not understand the situation at all. There is some behavioral data supporting this finding: The user interface offered a popup to review instructions or contact information. I tracked both and find that none of the participants ever opened these popups even though they were clearly visible in the decision screen's header and introduced in the instructions. To further analyze how participants coped with the

[16] The two-sided rank sum test (comparing differences between samples) yields a p-Value of 0.6059 for the mean Gini coefficient in last round of the game.

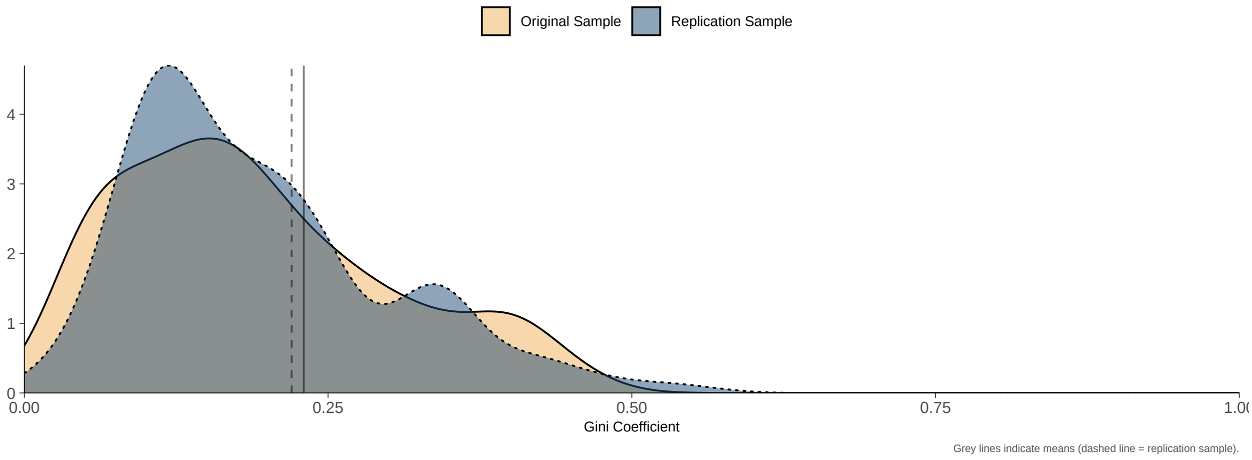

Figure 5: Groups' Gini coefficients (within groups) at the end of the game

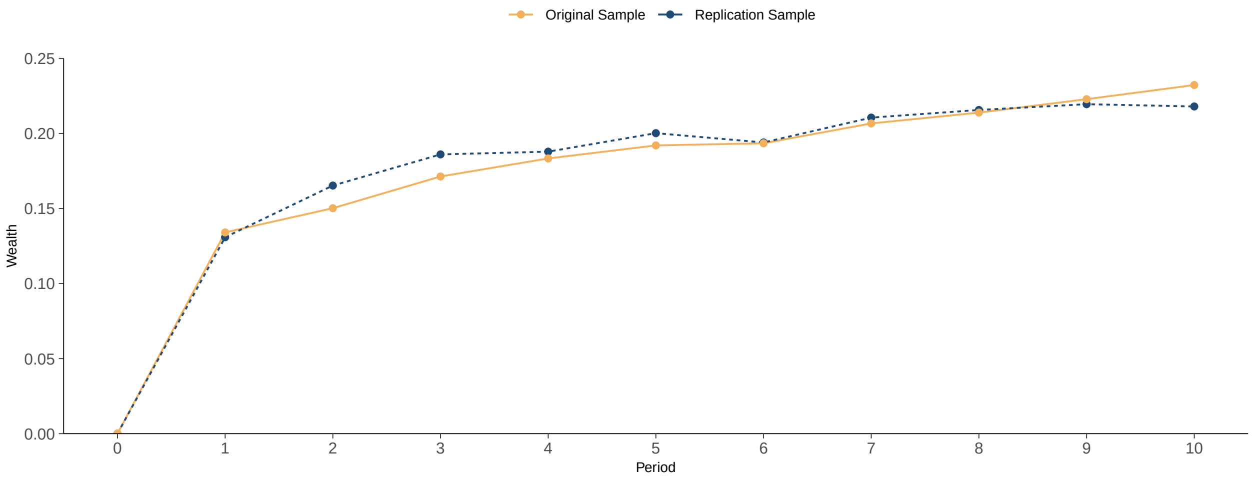

Figure 6: Average Gini coefficient (within groups) over time across samples

situation, I consider three additional metrics: selection into the experiment, attrition as well as the time spent on each page.

I first comment on the selection into the experiment: It was difficult to recruit the sample. The panel counted 1.209 non-students of which we were able to recruit 130 participants who finished the experiment—even though we varied the weekdays and timing of the sessions (which were conducted during a nationwide lockdown with home office regime). For this reason, we also recruited students in the last session which explains the relatively large number of showups in Table 2. Although I intended to refrain from the recruitment of students initially, this particular sub sample enabled me to investigate the generalizability of my results as I will discuss in Section 3.3. Alternatively, I also could have contacted a market research institute to recruit additional participants within a week. I refrained from doing so to contrast students and the general population sample, however.

Table 2: The Experimental Sessions' Meta Data

Session Code	Date	Time	Showups	Dropouts	Residuals	Participants	Observations
jyf8xd0s	2021-07-01	15:00	35	4	3	28	7
vggk2gh1	2021-07-03	13:00	20	8	0	12	3
8gi7c8xg	2021-07-09	13:00	21	5	4	12	3
d6jrsxnr	2021-07-23	14:00	75	8	3	64	16

Turning to the time spent on each page, I focus on the decision times in the dynamic public goods game as Anderhub et al. (2001) did. How many seconds did the participants need to make a decision in each period of the game? Not too many. Figure 7 illustrates an intuitive pattern: The first decision took about 22 seconds. The second decision–where participants first learned about the other group members' previous decisions–took longer (about 33 seconds). Subsequently, decision times first declined and stabilized at 19 seconds. Importantly, decision times were so short that crosstalk, that is, communication through private channels–a common concern[17] in online experiments–was unlikely, especially because it would require the identification of other group members.[18]

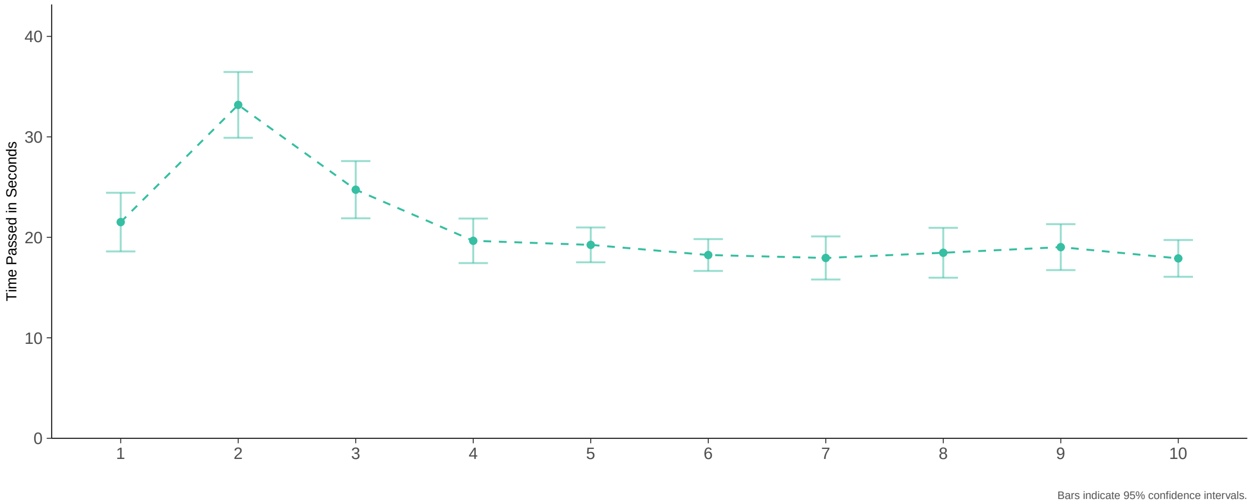

Figure 7: Average Time Spent for each Contribution per Period

[17] See, for instance, the discussion section in Arechar et al. (2018, p. 119).

[18] There were only 9 participants (from all four sessions) who needed more than 60 seconds to make the second decision.

Considering attrition, I find that it did not affect the interactive experiment at all. To elaborate, I differentiate between dropouts and residuals: Participants who could not be matched to other group members are called residuals. Participants who intentionally left the experiment are called dropouts. Residuals did not participate in the experiment *by design*. Dropouts did not participate in the experiment *by choice*. Out of 151 people who showed up, I count 10 residuals and 25 dropouts. All of the residuals waited to be matched to a group unsuccessfully before they got paid one Euro for their patience. In contrast, all of the dropouts left while reading the instructions and before being matched to other group members. Moreover, they got no payment at all. Hence, attrition was no concern considering the dynamic public goods game or the expenses.

Result 2. *Given the decision times and the fluent procedure, attrition was as negligible as it is in physical laboratories—where (a) not every invited person shows up and (b) a number of participants divisible by the group size is required as well.*

3.3. Generalizability

Goeschl et al. (2020) asked how much can we learn about voluntary climate action from the behavior in public goods games. Using a similar strategy, I answer the question for *dynamic* public goods games: *Not much*. Overall, there seems to be no association between choices in the voluntary climate action and the first period in the dynamic public goods game. Figure 8 shows a scatter plot of realized choices, with the percentage of endowment spent by each participant in the first period of the game on the x-axis and that spent in the VCA on the y-axis. In addition, the figure contains a fitted line of a linear model whose slope is indistinguishable from zero. No matter how much the participants contributed in the first round, they spent, on average, about 21% of their income on the VCA.

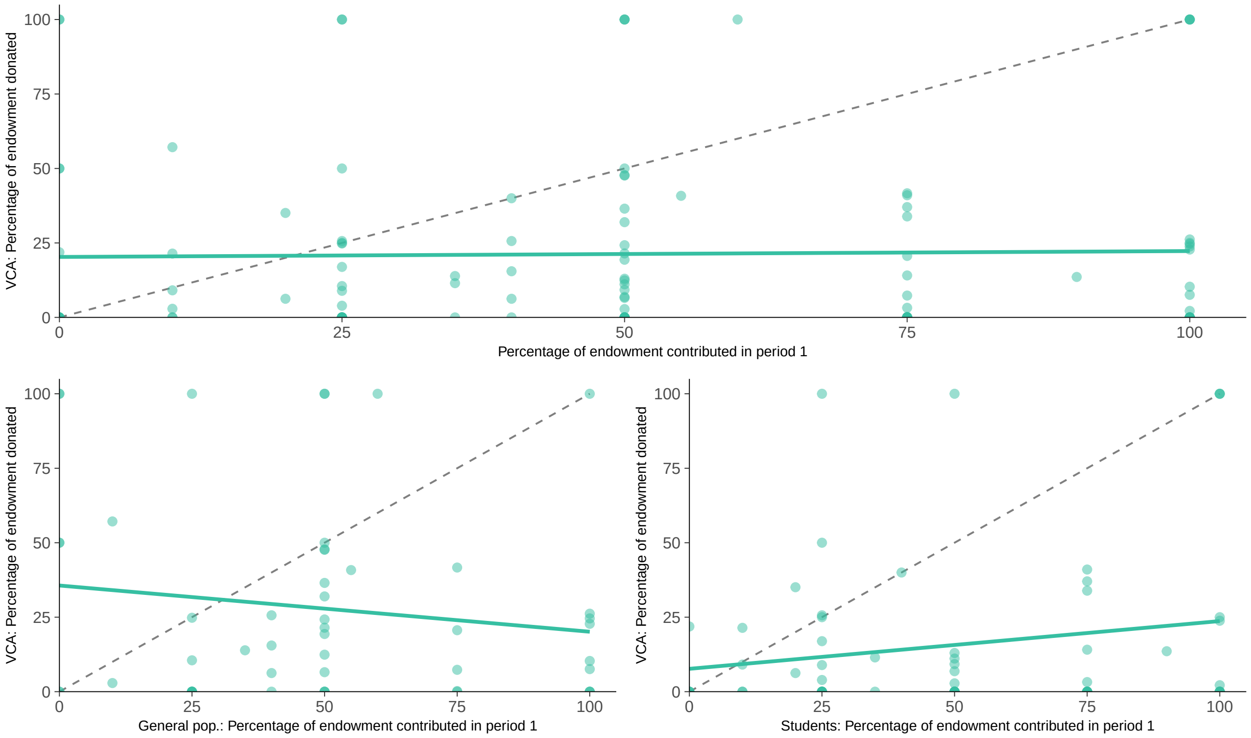

Figure 8: Scatter plot of average contributions in the dPGG and real giving task.

Does this result hold true if one zooms in and inspects the two samples separately? Yes. Even though the general population sample is a little more consistent than the student sample, both are contributing more in the abstract game than in the VCA. Figure 9 shows distributions of contributions across both choices for both samples. The left panels illustrate the behavior of the general population sample. The right panels illustrate the behavior of the general population sample. The top panels show the behavior in the VCA. The bottom panels show the behavior in first period of the game. A visual inspection shows that (a) both samples behaved similarly in the first period of the game but (b) different ($p = 0.03$) in the VCA. Furthermore, (c) the behavior in the first period of the experiment predicts the general population sample's behavior in the VCA worse than student sample's behavior ($p = 0.07$). Finally, (4) contributions in the abstract public good game are higher than contributions to the real public good of climate change mitigation.

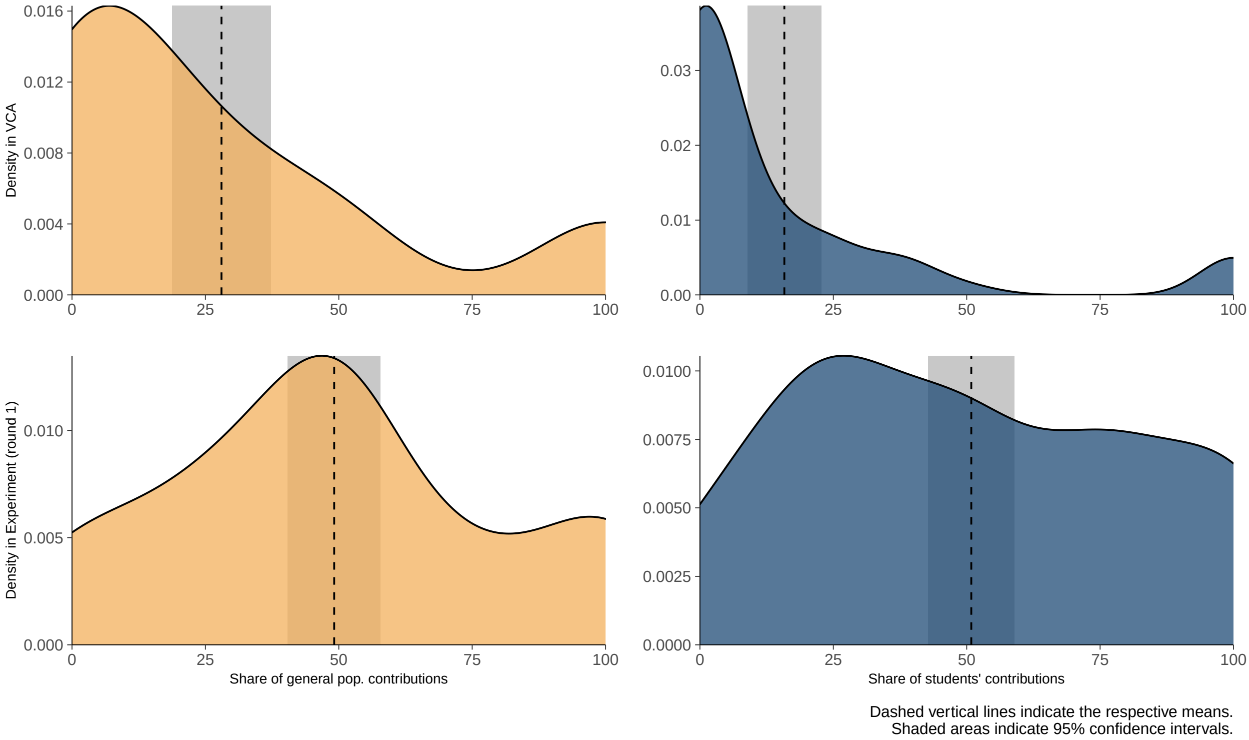

Figure 9: Kernel distributions of contributions across tasks and subject pools.

Result 3. *Overall, contribution behavior is uncorrelated with the willingness to contribute to real public goods. This holds true for more representative samples and—to a lower degree—for student samples.*

4. Conclusion

The goal of the experiment was to replicate specific experiments of GMTV in an online setting using a general population sample. The results suggest that it is important to replicate experiments—both purely and scientifically (Hamermesh, 2007, p. 716)—before drawing conclusions about generalizability.

The three most important findings are as follows: First, the contribution behavior in my experiment is statistically similar to the behavior reported in the original study. Consequently, the outcomes growth and inequality are similar as well. Second, the online experiment proceeded fluently such that dropouts were no concern. Third, contribution behavior in my abstract setting is, if anything, only weakly linked to behavior in the real world.

The significance of the first result is that similar procedures led to replicable findings under different circumstances across two different samples. The second result is of methodological importance: It highlights that even logistically complex experiments can be conducted online with—not only with clickworkers but also with a true general population sample. The third result questions whether recruiting from more representative samples is worth the efforts because it decreases transferability of results to the real world—at least in this specific case.

5. A: Pure Replication

This section comments on two errors as well as a misconception I found in the original data.[19] Before I proceed to explain this in more detail I would like to say that the results of the original paper still hold after the error is fixed and that the authors responded kindly and quickly, showing an interest in solving the issue. In fact, some explanations in this section stem from input provided by the authors.

5.0.1. Error 1: The Gini coefficient

The Gini coefficient is wrongly computed in some periods for some group members. The authors found that this happened whenever two group members had exactly the same endowment because the program failed to rank these group members for further calculations.

Table 3 illustrates this problem. It shows group 101 in period 5 and documents that the Gini coefficient differs among group members. According to the authors, the Gini coefficient should equal `GINI`=0.127 for all subjects in the group. Instead, participant `112` and `113` who have an equal endowment deviate from that value. Importantly, the `DescTools::Gini()` function in the statistical software `R` does not yield this error, which is why I use that function for my calculations using both my as well as the original data.

Table 3: Subset of Data illustrating the Gini Coefficient's Error

exp_num	gr_id	per	subj_id	tokens	other1	other2	other3	gini	GINI
1	101	5	111	42	27	27	30	0.127	0.127
1	101	5	112	27	42	27	30	0.111	0.127
1	101	5	113	27	42	27	30	0.111	0.127
1	101	5	114	30	42	27	27	0.127	0.127

5.0.2. Error 2: The share of endowments contributed

The original data provides a wrong measure of the share of endowments contributed (`mean`) because it relies on a lagged endowment (`gdp`). More precisely, the authors used the following STATA code for their calculations:

```
*tsset subj_id per
*gen mean=sum/l.gdp
```

Table 4 reports participant 111 in group 101 in experiment 1 over three periods. Both the `gdp` (that is, the sum of the group's endowments at the beginning of the period) as well as the `sum` (that is, the sum of the group's contributions) are group-level variables.

Table 4: Subset of Data illustrating the Means's Error

exp_num	gr_id	per	subj_id	gdp	sum	mean	MEAN
1	101	4	111	116	18	0.168	0.155
1	101	5	111	126	18	0.155	0.143
1	101	6	111	136	17	0.135	0.125

Calculating the share as `MEAN=sum/gdp` solves the problem and yields $\frac{18}{126} = 0.143$ in period 5. I thus, used this proposed definition for all my calculations using both my as well as the original data.

[19] The data can be found in the supplementary materials they provide in their online appendix.

5.0.3. The misconception: Timing

The authors wrote a note stating that the Gini coefficient as well as the wealth in the paper always refer to the situation at the start of a period and that they clarify this because the paper (last paragraph at the bottom of page 5), says that wealth is defined as the endowment at the beginning of the following period. Furthermore, they write that this error came about as they switched between these two definitions during the course of revising the paper.

I argue that it makes more sense to calculate the variables as they state in the paper. More precisely, I think that the wealth at the *beginning* of a period is less interesting than the wealth at the *end* of a period for two reasons: First, there is no need for such a variable because it already exists (the endowment). Second, this definition yields a value that is determined by the design of the game but misses an important outcome at the end of the game. To illustrate this, note that the wealth would be defined as four times the initial endowment in period 1. Also note that the very last value would equal the wealth at the beginning of the last period and says nothing about the outcome of that period. Because the contributions often drop in the last period, this outcome is of particular interest (yet, not represented in the data). Moreover, this definition of wealth yields more informative values to calculate the Gini coefficient for the same reasons: We know that the Gini coefficient is zero *before* the participants made any decision by design. We do no know the inequality at the very end of the game—and the current definition does not tell us.

For these reasons, I define wealth and inequality measures as the outcomes of a period for all of my calculations using both my as well as the original data.[20]

[20] Accordingly, the definition of `GINI` I provide in Table 3 is not the definition I used to calculate the current period's Gini coefficient but the previous period's Gini coefficient.

References

Anderhub, V., Müller, R., Schmidt, C., 2001. Design and evaluation of an economic experiment via the internet. Journal of Economic Behavior and Organization 46, 227–247. URL: https://www.sciencedirect.com/science/article/pii/S0167268101001950, doi:https://doi.org/10.1016/S0167-2681(01)00195-0.

Arechar, A.A., Gächter, S., Molleman, L., 2018. Conducting interactive experiments online. Experimental economics 21, 99–131. URL: https://link.springer.com/article/10.1007/s10683-017-9527-2, doi:https://doi.org/10.1007/s10683-017-9527-2.

Berlemann, M., Roggenkamp, H., Traub, S., 2021. Replication: Growth and inequality in public good provision (No-Punish-10) by Gächter et al. (2017). Technical Report. doi:https://doi.org/10.1257/rct.7902-2.0.

Bock, O., Baetge, I., Nicklisch, A., 2014. hroot: Hamburg registration and organization online tool. European Economic Review 71, 117–120. URL: https://www.sciencedirect.com/science/article/pii/S0014292114001159, doi:https://doi.org/10.1016/j.euroecorev.2014.07.003.

Chen, D.L., Schonger, M., Wickens, C., 2016. otree-an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance 9, 88 – 97. URL: http://www.sciencedirect.com/science/article/pii/S2214635016000101, doi:10.1016/j.jbef.2015.12.001.

Eichenseer, M., Moser, J., 2019. Leadership in dynamic public good provision: Endogenous growth and inequality. Available at SSRN 3269341 .

Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90, 980–994. URL: https://www.aeaweb.org/articles?id=10.1257/aer.90.4.980, doi:10.1257/aer.90.4.980.

Goeschl, T., Kettner, S.E., Lohse, J., Schwieren, C., 2020. How much can we learn about voluntary climate action from behavior in public goods games? Ecological Economics 171, 106591. URL: https://www.sciencedirect.com/science/article/pii/S0921800919302745, doi:https://doi.org/10.1016/j.ecolecon.2020.106591.

Goodman, J.K., Paolacci, G., 2017. Crowdsourcing Consumer Research. Journal of Consumer Research 44, 196–210. URL: https://doi.org/10.1093/jcr/ucx047, doi:10.1093/jcr/ucx047, arXiv:https://academic.oup.com/jcr/article-pdf/44/1/196/25496127/ucx047.pdf.

Gächter, S., Mengel, F., Tsakas, E., Vostroknutov, A., 2017. Growth and inequality in public good provision. Journal of Public Economics 150, 1–13. URL: https://www.sciencedirect.com/science/article/pii/S0047272717300361, doi:10.1016/j.jpubeco.2017.03.002.

Hamermesh, D.S., 2007. Viewpoint: Replication in economics. Canadian Journal of Economics/Revue canadienne d'économique 40, 715–733. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2966.2007.00428.x, doi:https://doi.org/10.1111/j.1365-2966.2007.00428.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2966.2007.00428.x.

Holt, C.A., Laury, S.K., 2002. Risk aversion and incentive effects. American Economic Review 92, 1644–1655. URL: https://www.aeaweb.org/articles?id=10.1257/000282802762024700, doi:10.1257/000282802762024700.

Parsons, S., Azevedo, F., Elsherif, M.M., Guay, S., Shahim, O.N., Govaart, G.H., Norris, E., O'mahony, A., Parker, A.J., Todorovic, A., et al., 2022. A community-sourced glossary of open scholarship terms. Nature human behaviour 6, 312–318. URL: https://doi.org/10.1038/s41562-021-01269-4, doi:10.1038/s41562-021-01269-4.