

# Revisiting ‘Growth and Inequality in Public Good Provision’ —Reproducing and Generalizing Through Inconvenient Online Experimentation

Hauke Roggenkamp<sup>a,b</sup>

<sup>a</sup>*Helmut Schmidt University, Holstenhofweg 85, Hamburg, 22043,*

<sup>b</sup>*University of St. Gallen, Torstrasse 25, St. Gallen, 9000,*

---

## Abstract

I revisit the dynamic public goods game developed by Gächter et al. (2017) to study cooperation under dynamic interdependencies. Collecting data from both a convenient (students) and an inconvenient (general population) sample, I not only reproduce some of the authors’ original observations but also test their novel game’s generalizability. Appending a charitable dictator game, I find no correlations between behavior in the charitable context and the dynamic game. This applies to students and the general population sample alike. Because the study of inexperienced general population samples raises methodological challenges, such as fatigue and dropouts, this research approaches them. In doing so, I provide simple solutions to run reliable interactive experiments online. Furthermore, this article showcases the use of literate programming and version control which I argue are convenient tools to make pre-registrations more credible and flexible.

*Keywords:* Public goods, Dynamic games, Reproducibility, Generalizability, Online Experiment, Climate change mitigation

---

## 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 environmentally friendly policies today not only reduces carbon dioxide emissions 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 requires more effort in the future compared to a path that includes immediate action (Hänsel et al., 2022). Aiming at certain goals, today’s actions (or the omission thereof) not only affect intermediate outcomes but also the number of paths one can choose from that lead to that specific goal.

---

\*Corresponding author

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

Public good (or public bad) games—although often intended to inform climate policies (e.g. [Milinski et al., 2006](#); [Tavoni et al., 2011](#); [Hauser et al., 2014](#); [Brick and Visser, 2015](#); [Vicens et al., 2018](#); [Calzolari et al., 2018](#); [Cook et al., 2019](#))—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 action space in subsequent periods. To see the lack of realism, consider carbon dioxide emissions, where the current stock will last for well over a millennium ([Inman, 2008](#); [Calzolari et al., 2018](#)). Playing with fresh endowments in each period is as if one could just undo carbon dioxide emissions at no cost.

A game designed by Gächter, Mengel, Tsakas & Vostroknutov ([2017](#), hereafter, GMTV) as well as Stefan Große (unpublished), shows that it is fairly simple to add this element of realism. They incorporate interdependencies into a *dynamic* public goods game (dPGG) by defining endowments as the income of previous periods. Consequently, participants’ actions in a given period affect their number of actions in subsequent periods: the more (less) they earn now, the more (less) they can contribute in the next period. Importantly, this modification qualitatively yields the same rational predictions as the static game (i.e. free-riding and the under-provision of the public good). It is thus, equally well suited to study dilemma situations.

Because there is surprisingly little experimental research on interdependencies<sup>1</sup>, I reproduced one of GMTVs’ treatments to compare dynamics across (in)experienced samples to investigate its generalizability. [Goeschl et al. \(2020\)](#) find that static public goods games do not generalize well to real-world climate action. They also find that generalizability depends on the structural resemblance of the public goods game with the context of climate change mitigation: Greater resemblance improves generalizability. Because GMTVs’ dynamic setting has a more realistic property—namely, interdependencies—one would expect it to be better suited to inform public policy. To test this intuition, I not only ran the experiment with different samples but also observed the participants’ behavior in voluntary climate actions (VCA).<sup>2</sup> This yielded a setting similar to the one of [Goeschl et al. \(2020\)](#), which allows me to analyze how behavior in the abstract game translates into real-world action across samples. This research shows that the dynamic setting *does not* add any advances to the generalizability of results.

[Arechar et al. \(2018\)](#) conducted static public goods games in the lab and on MTurk to draw (and report) lessons from online experimentation. This study extends this literature (see, e.g., [Krantz et al., 1997](#), [Amir et al. \(2012\)](#); [Klein et al., 2014](#); [Paolacci and Chandler, 2014](#); [Goodman and Paolacci, 2017](#); [Snowberg and Yariv, 2021](#); [Gupta et al., 2021](#); [Buso et al., 2021](#)) by focusing on an *inconvenient* sample (i.e. one that is a

---

<sup>1</sup>See e.g. [Battaglini et al. \(2016\)](#), [Rockenbach and Wolff \(2017\)](#) or [Eichenseer and Moser \(2019\)](#).

<sup>2</sup>The VCA is a (charitable) dictator game where each participant is a dictator dividing her budget between herself and some organization linked to the reduction of CO2 emissions. [Eckel and Grossman \(1996\)](#) were the first to implement a charitable dictator game observing contributions of 30% of the endowments. Like [Goeschl et al. \(2020\)](#), [Carpenter et al. \(2008\)](#) report that students make lower contributions to charity than community members.

completely inexperienced sample (see also [Benndorf et al., 2017](#)) that has not been exposed to interactive experiments before) playing a computationally more complex game during a time that was characterized by more inattentiveness in online samples ([Arechar and Rand, 2021](#); [Peyton et al., 2022](#)). This required me to design robust (and thus, more complex) software to minimize attrition. I collected para data ([Parsons et al., 2022](#), p. 12)—which capture screen time and survey navigation, for instance—to assess the desired fluency and feasibility of the experiment. Like [Arechar et al. \(2018\)](#), this research is of practical relevance as it reports on the robust design that made the dynamic game feasible.

Taken together, this study makes three contributions. First, it reproduces parts of GMTVs’ 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, this paper supports critics arguing that findings from abstract games do not generalize well—not even with a more representative sample.

After commenting on transparent research practices and reporting the methods in [Section 2](#) and [Section 3](#), this paper is organized along these findings: The confirmatory reproduction is reported in [Section 4.1](#), the online feasibility in [Section 4.2](#), and the generalizability in [Section 4.3](#). [Section 5](#) concludes. Because I observed some (qualitatively inconsequential) differences in reproducing GMTVs’ analysis using their data, I report these in the [Appendix](#).

## 2. Transparency

Well before credibility crises comprised a variety of disciplines<sup>3</sup>, [Bem \(1987, p. 2\)](#) indirectly conceded discrepancies between fuzzy research processes and the polished article that results from it:

*There are two possible articles you can write: (a) the article you planned to write when you designed your study or (b) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (b).*

Now, 20 years later, we accustom ourselves to mechanisms designed to unravel the fuzzy back-and-forth between exploratory and confirmatory research. Prominently, pre-registrations and pre-analysis plans (PAP) (see, e.g., [Brodeur et al., 2022](#)) were established to tie the researchers’ hands with respect to p-hacking and ex-post theorizing when doing confirmatory research—reducing observable discrepancies between Daryl Bem’s above-mentioned (a) and (b).

---

<sup>3</sup>See e.g., [Camerer et al. \(2016\)](#); [Brodeur et al. \(2016\)](#); [Brodeur et al. \(2020\)](#); [Christensen and Miguel \(2018\)](#); [Ferraro and Shukla \(2020\)](#); [Page et al. \(2021\)](#) who focus on economics.

While pre-registrations and PAPs could help to make empirical sciences more credible, they come at costs that are discussed by [Olken \(2015\)](#), [Page et al. \(2021\)](#), as well as [Krishna \(2021\)](#), [Simmons et al. \(2021\)](#), and [Pham and Oh \(2021\)](#). One important concern is that strict pre-registrations may severely discount any additional analysis of the data inspired by surprise results. In practice, there are at least two ways to approach that problem: First, one could (and should) add unforeseen analyses to the article but label them as exploratory. Second, one could write a minimal pre-registration leaving enough degrees of freedom for the researcher to incorporate such surprises without further notice. The key difference between the two approaches is the deliberate omission of transparency.

For the field of neuropsychopharmacology [Waldron and Allen \(2022\)](#) document that the depth of pre-registrations differs indeed: it is spanning from notes on hypotheses only to fully specified, detailed experimental protocols with accompanying power analyses and sampling plans. Treating all pre-registrations equally may thus lead to an adverse effect: If one can acquire the ‘*pre-registered*’ label with minimal effort and maximum flexibility, the label itself may run into danger to become worthless and to miss the target of making science more credible.

This erosion of credibility can be avoided by the means of transparency—not because transparency can tighten the ties but because it facilitates scrutiny, policing (see [Ankel-Peters et al., 2023](#)) and forensics (see, e.g., [Simonsohn et al., 2023](#)): transparent research practices make it easier for others to track and understand a researcher’s reaction to surprises in the data, which, in turn, gives the researcher more flexibility to react to these surprises. As additional transparency is associated with additional effort, I propose version control as a natural and easy-to-implement solution augmenting pre-registrations.

Contemporary experimental research is, to a certain degree, a computational science: experiments often require complex software and the analysis thereof can result in complicated scripts. In addition, much research is carried out in collaborative efforts. For these reasons alone, it makes sense to implement version control systems such as GitHub and GitLab as well as OSF (albeit less powerful) that ensure that files are stored consistently and that changes are tracked. When these systems are implemented in the whole research process, one can archive different states of the software and visualize differences between different states. For instance, one can archive the analysis before and after the data was collected or the experimental software used in a pilot and the eventual experiment. Importantly, one can keep projects private and change their visibility to public only after the corresponding article is accepted for publication. Taken together, these features *automatically* enhance transparency and facilitate scrutiny, policing, and forensics whilst also making software development and collaboration more efficient. As a consequence, the implementation and eventual disclosure of version control come not only with meta-scientific but also with organizational benefits.

The experiment reported in this article qualifies for the ‘*pre-registered*’ label at first glance: The analysis

was pre-registered in the American Economic Association’s RCT Registry (Berlemann et al., 2021). Further, I pre-registered the exact analyses I planned to run (PAP) when I designed the experiment on GitHub.<sup>4</sup>

At second glance, the analysis that made the most sense after having collected the data and that is reported here is a little different though: Because the more representative subject pool was exhausted earlier than expected, I recruited students which opened up a new research direction: assessing generalizability. Hence, this article combines both confirmatory as well as exploratory research.

Accordingly, I edited many of the analysis scripts.<sup>5</sup> However, because the originally planned analysis code is archived and reproducible, these changes are transparent. Importantly, I followed a literate programming approach (Knuth, 1984; Akhtar and Ye, 2023). Hence, all documents needed to analyze the data and to write the report stem from the same source. This establishes consistence between the commands one tells the statistical software to do and the explanation one tells human beings one told the statistical software to do. As such, (a), (b), and the transition from (a) to (b) are not only transparent but also comprehensible.

### 3. Methodology

In the terminology of Hamermesh (2007), I ran both a *pure* as well as a *scientific* reproduction<sup>6</sup> of one treatment of GMTVs’ dynamic public goods game. The pure reproduction re-analyzes the original data. Appendix A documents the errors I identified in the original paper. The scientific reproduction, where I utilize a different sample drawn from a different population in a different situation, is described in the following sections.

#### 3.1. Experimental Design

The design builds on the workhorse model Zelmer (2003, p.301) describes in her meta-analysis, where

*subjects are divided into groups and play the same game for a finite number of periods. Each period, every subject is endowed with an income [...] The subject must then divide this income between a contribution to a private account [...] that yields a constant return to themselves only and a contribution to a public account [...] where consumption benefits accrue to all group members. At the end of each period, subjects typically learn the aggregate contribution to the public good by all members of their group and their earnings for the period.*

---

<sup>4</sup>[bit.ly/3Pnak5H](https://bit.ly/3Pnak5H). To run the code, you need to execute the .Rmd files in this repository in the order that is indicated by file names. Detailed instructions can be found in the README file accompanying the analysis scripts.

<sup>5</sup>The source code of this document contains the analysis and can be found [here](https://bit.ly/43LCyLw): <https://bit.ly/43LCyLw>

<sup>6</sup>In fact, Hamermesh (2007) uses the term *replication* instead of *reproduction*. The reason I deviate is that I do not compare different treatments with each other to *replicate* a treatment effect. Instead, I *reproduce* the results of one specific treatment arm.

Considering the design, the ‘only’ differences between such a *static* game (with partner matching) and GMTVs’ design are *dynamics*: Instead of receiving fresh endowments every period, participants receive one endowment only at the beginning of the first period. A participant’s endowment in the second period is the wealth she accumulated in the first period. A participant’s endowment in the third period is the wealth 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 goods game.

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<sup>7</sup>,  $T = 10$  periods, a private account with a return of 1 and a group account with a return of 1.5 ( $\Rightarrow \text{MPCR} \equiv \frac{1.5}{4}$ ). With  $i$ ’s contribution in period  $t$  being  $c_i^t$ , the model looks as follows:

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

### 3.2. Voluntary Climate Action (VCA)

Like GMTV I employ a real giving task after the abstract game. Specifically, I employed a (charitable) dictator game where each participant is a dictator dividing her budget between herself and some organization. Whereas GMTV chose *Doctors without Borders* as a receiver, I chose another organization in another context: like [Goeschl et al. \(2020\)](#), I employed a VCA, where participants could donate any amount of their earnings to offset carbon dioxide (that is, retire emission permits from the EU ETS).<sup>8</sup> 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 was to be posted immediately after the conclusion of the sessions 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.

### 3.3. Recruitment and Sample Characteristics

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

---

<sup>7</sup>A token was worth 0.05 Euros.

<sup>8</sup>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 GMTVs’ procedure.

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 single-player experiments. Recruiting them, I excluded those who have participated in interactive experiments such as public goods games before, to keep the two distinct subject pools comparable in terms of their experience. 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 the results of my experiment with the results of GMTVs’ 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 of Hamburg’s students (both in July 2021). Table 1 gives an overview.

Table 1: Sample Properties

sample	Hamburg Citizens			Hamburg Students			Nottingham Students			
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD	Test
Age	52	48	15	64	26	7.6	92	33	2.6	F= 88.98***
Female	52	0.48	0.5	64	0.56	0.5	92	0.38	0.49	F= 2.596*
Risk Aversion (1-12)	20	6.6	2.6	58	6	2	87	6.4	1.3	F= 1.406

Statistical significance markers: \*  $p < 0.1$ ; \*\*  $p < 0.05$ ; \*\*\*  $p < 0.01$

Overall, I recruited 116 participants for the experiment. The three samples differ significantly with respect to their age. The non-student sample is more diverse compared to the two student samples.

### 3.4. 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 in implementing the experiment.<sup>9</sup>

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

<sup>9</sup>see Zhou and Fishbach (2016) for a nice illustration of how (selective) attrition affects identification.

user interface interactive, appealing, and easy to navigate. The [Highcharts library](#) made it easy to visualize results and communicate dynamics. Insofar, oTree served as a good tool to enhance the participants' user experience and thus, to make dropouts less likely.

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 4.2 shows that the first feature became effective in some cases, whereas the second feature did not.

### 3.5. 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 to avoid confusion in later stages ([Ferraro and Vossler, 2010](#)). 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 the results of all periods. Subsequently, they made their VCA decision, before I elicited risk preferences ([Holt and Laury, 2002](#)) and finished with GMTVs' questionnaire.

While I stuck to GMTVs' 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) which also affected the graphical user interface participants were exposed to. Third, GMTV gave participants the opportunity to donate to *Doctors without Borders* whereas I offered carbon dioxide offsets.

The experiment lasted around 25 minutes on average. The earnings averaged 11.23 Euros ( $sd = 4.85$ ).<sup>10</sup>

---

<sup>10</sup>This value include earnings from the incentivized risk elicitation task that is not part of the analysis.



## 4. Results

### 4.1. Reproducibility

Throughout this section, I pool data from both the citizens as well as the students from Hamburg and compare it with the students GMTV recruited in Nottingham. I refer to the two samples as the *reproduction sample* ( $N_r = 116$ ) and the *original sample* ( $N_o = 92$ ), respectively.

#### 4.1.1. Contribution Behavior

First, I ask whether the samples differ with respect to their initial contributions to the public good. Is the reproduction sample (consisting of both students and non-students) more pro-social than the original sample? A two-sided rank sum test reveals that it is not ( $p=0.393$ ). Both samples contributed 10 tokens, that is, 50% of their endowments on average (median and mean). Moreover, both samples' initial contributions resemble the initial contributions participants usually make in the static game with partner matching.<sup>11</sup> However, in the dynamic game presented here, I am particularly interested in the subsequent periods because differences add up exponentially. Do the two groups remain similar over the course of time?

In particular, do the two samples' contributions follow the same path over the 10 periods they played? The answer is *yes and no*. Figure 1 and Table 6 illustrate 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 more than the reproduction sample *in all but the first and last period*.<sup>12</sup> For this reason, the original sample's behavior differs from the reproduction sample's behavior in two aspects: it contributes more and exhibits a considerable drop in the last period (whereas the reproduction sample's contributions flatten).

Note that increasing contributions over time imply increasing endowments over time. Hence, absolute contributions do not tell much about the willingness to cooperate. For this reason, the right panel in Figure 1 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 reproduction sample's share of contributions declines faster. This is mirrored by a two-sided rank sum test which is only significant at a five percent level in periods three and four.

Again, both samples' behavior resembles the contributions participants usually make in the static game with partner matching: contributions equal approximately half of the endowments in the very first period

---

<sup>11</sup>See Figure 3B in Fehr and Gächter (2000, p.989), for instance.

<sup>12</sup>In period five, this difference is significant: a two-sided rank sum test yields  $p=0.0486$ .

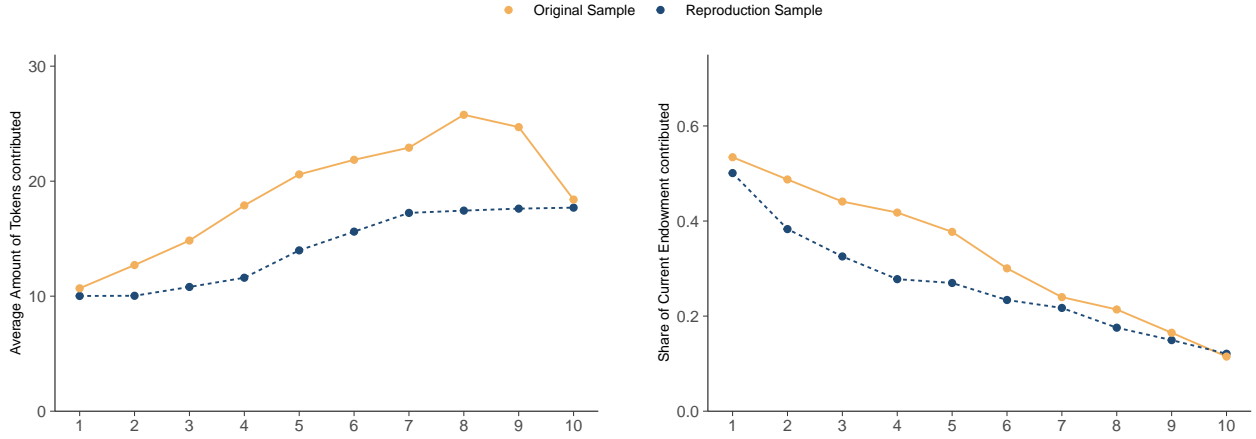


Figure 1: The average amount of tokens contributed over time across samples.

and decrease to around ten percent of endowments by the last period.<sup>13</sup> In the dynamic game presented here, however, different paths lead to different levels of wealth – even if they share the same start- and endpoints. I am thus, more interested in the contributions’ implications for wealth generation and growth.

#### 4.1.2. Wealth Creation

How do the different contribution paths translate into wealth?<sup>14</sup> Given that the original sample contributed more, one would expect the respective groups to be more wealthy. A mere mean comparison indicates just that: An average group in the original sample accumulated about 478 tokens. In contrast, the average group in the reproduction sample accumulated about 380 tokens. This difference is insignificant at conventional levels though: A two-sided rank sum test (comparing differences between samples) yields  $p=0.136$  for the mean stock in the last period of the game.

Although there clearly is growth, groups do not realize the maximal potential efficiency: under full cooperation, a group can accumulate a maximum of 4613 tokens or EUR 230. This is depicted in the left panel of Figure 2, 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. Despite somewhat differing contribution behavior between samples, 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*

<sup>13</sup>The right panel is thus, comparable to the visualizations *and results* in the static game. See, for instance, Figure 1B in Fehr and Gächter (2000, p.986).

<sup>14</sup>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).

a significant difference. The right panel of Figure 2 depicts heterogeneity: In the reproduction sample, the richest group earned 1425 tokens (which is about 1781% of the initial endowment) whereas the poorest group ends up with 92 tokens (115%). More broadly, the reproduction 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 reproduction sample was drawn from a more heterogeneous (non-convenience sample). Does it differ within groups?

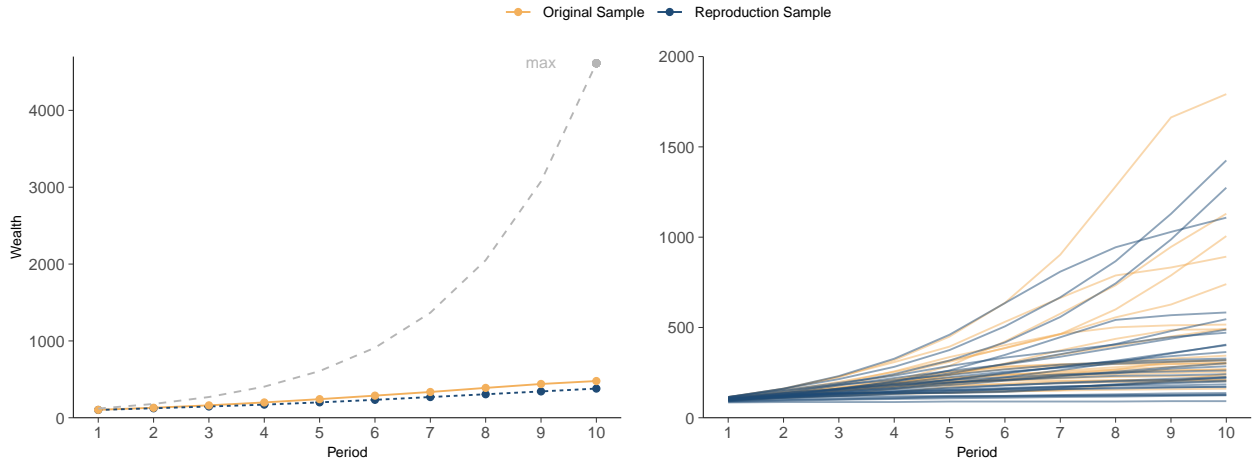


Figure 2: Average wealth over time across samples.

#### 4.1.3. 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 3 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.<sup>15</sup> 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 3 also 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.<sup>16</sup>

**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.*

<sup>15</sup>The two-sided rank sum test (comparing differences between samples) yields  $p=0.606$  for the mean Gini coefficient in the last round of the game.

<sup>16</sup>In each and every period, the two-sided rank sum test comparing gini coefficients between both sample yields  $p$ -values way over ten percent.

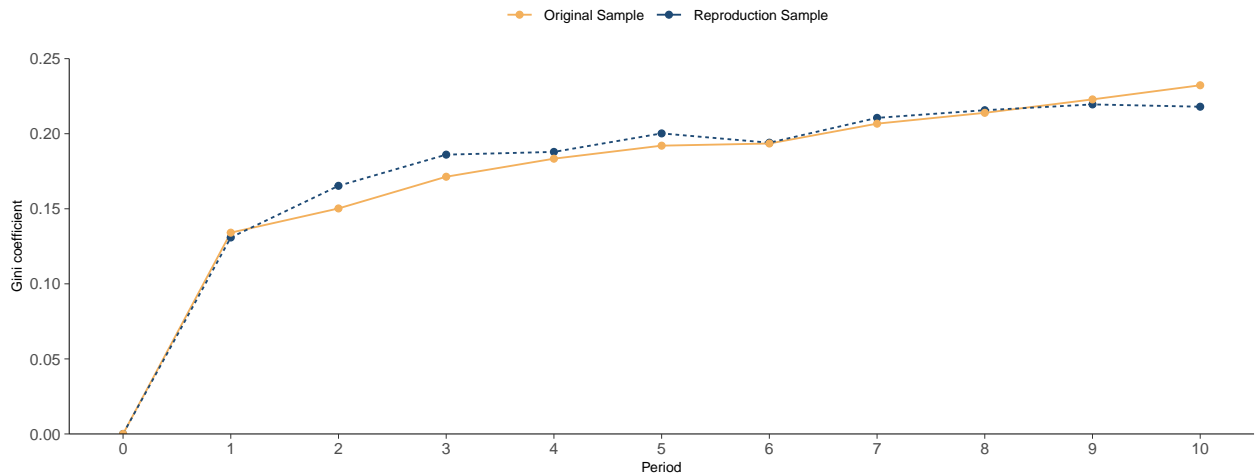


Figure 3: Average Gini coefficient (within groups) over time across samples

This is remarkable given the different sample and language, the different software and user interface as well as the online setting during the COVID-19 pandemic. The result suggests that by and large, the sum of these factors did not affect people’s preferences towards cooperation.

#### 4.2. Online Feasibility

Throughout this section, I do not consider GMTVs’ data and pool data I gathered from the citizens as well as the students from Hamburg ( $N = 151$ ).

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. Some behavioral data are 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 screens’ header and introduced in the instructions. To further analyze how participants coped with the 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 I was able to recruit 130 participants who finished the experiment—even though I varied the weekdays and timing of the sessions (which were conducted during a nationwide lockdown with home office regime). For this reason, I 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

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

students initially, this particular sub-sample enabled me to investigate the generalizability of my results as I will discuss in Section 4.3.

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 decide in each period of the game? Not too many. Figure 4 illustrates a plausible 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<sup>17</sup> in online experiments—was unlikely, especially because it would require the identification of other group members.<sup>18</sup>

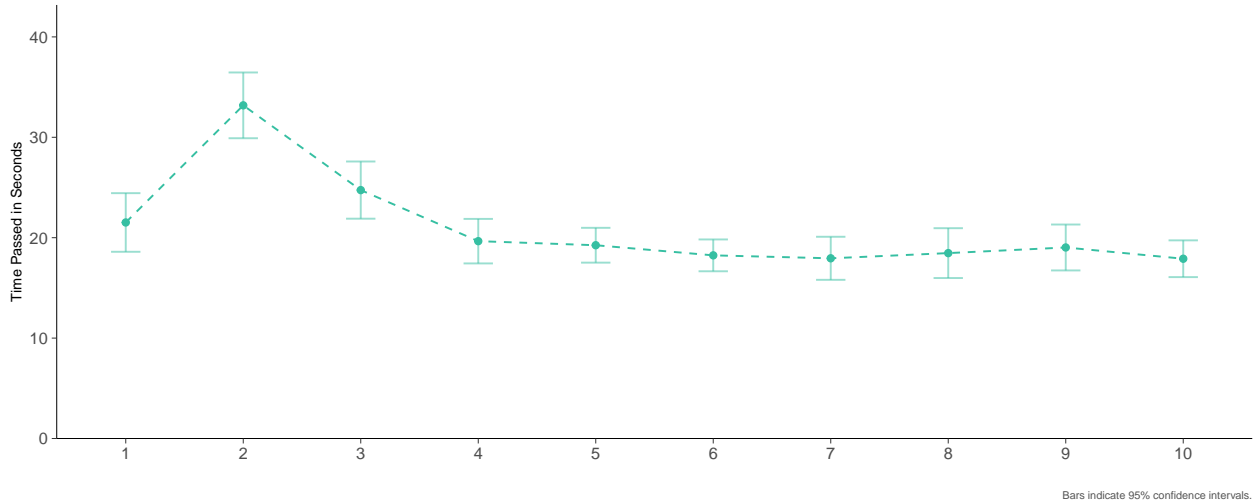


Figure 4: Average Time Spent for Each Contribution per Period

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

<sup>17</sup>See, for instance, the discussion section in [Arechar et al. \(2018, p. 119\)](#).

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

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) several participants divisible by the group size are required as well.*

#### 4.3. Generalizability

As before, I do not consider GMTVs' data in this section. Instead, I will differentiate between data from the citizens and the students from Hamburg. I refer to the two samples as the *students* ( $N_s = 64$ ) and the *general population sample* ( $N_g = 52$ ), respectively.

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 a *dynamic* public goods game: *Not too 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 5 shows distributions of contributions across both choices for both samples. The top panels illustrate the behavior of the general population sample. The bottom panels illustrate the behavior of the student sample. The left panels show the behavior in the VCA. The right panels show the behavior in the first period of the game. A visual inspection shows that (a) mean contributions are positive in both tasks for both samples.<sup>19</sup> (b) Furthermore, average contributions are lower in the VCA. (c) In contrast to the observation of Goeschl et al. (2020), I do not observe a difference between samples in the abstract game's contribution behavior. (d) However, the student's share of income contributed to the VCA is significantly lower than the general population sample's contributions (two-sided rank sum test,  $p=0.009$ ). Taken together, these aggregate results indicate, that the consistency between tasks is higher for the general population than it is for students. Or, to put it differently, the general population's behavior in the abstract game better predicts their behavior in a real-world mitigation context.

Table 3 shows that this is not the case. It reports tobit regression results cautioning against transferability from dPGG results to real-world mitigation behavior: The share of endowment contributed in the first

---

<sup>19</sup>The same holds true for median contributions.

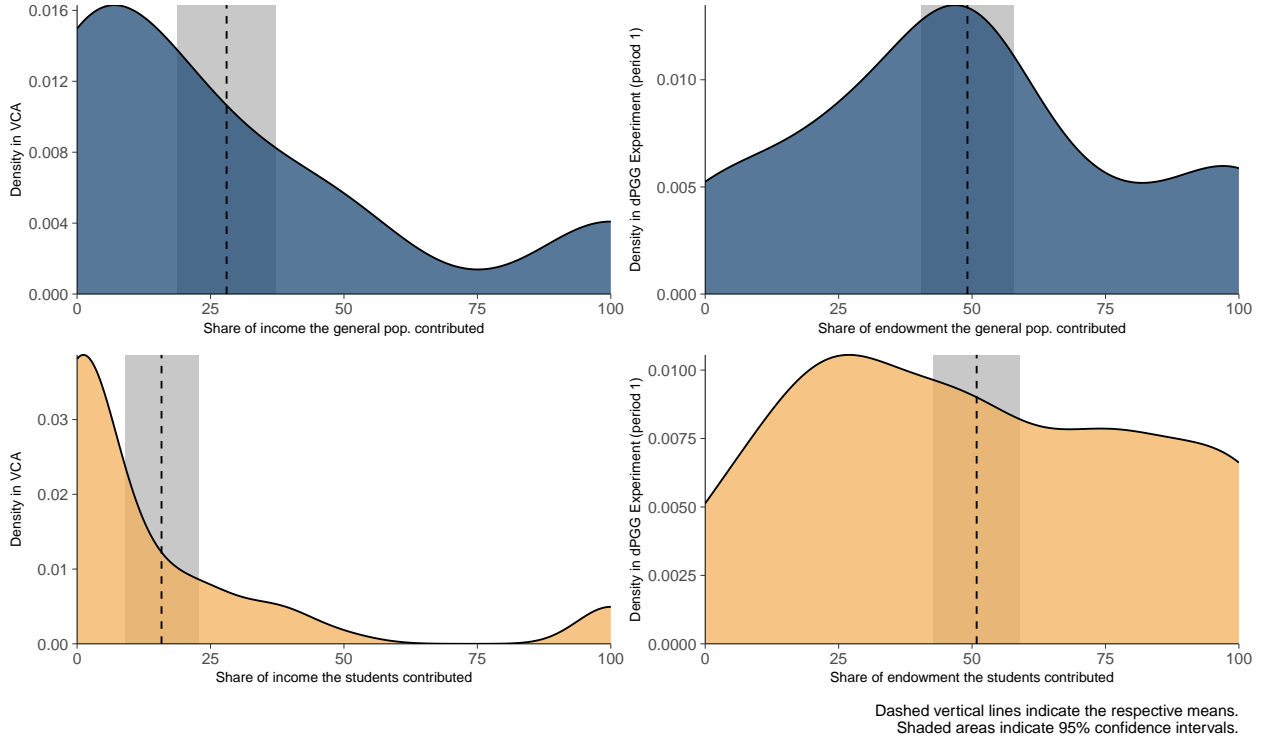


Figure 5: Kernel distributions of contributions across tasks and subject pools.

period (displayed in the first row) does not predict the share of earnings donated as a VCA. The student status negatively affects VCA donations in column two but disappears if one controls for age in column three. Importantly, the interaction between student status and first-period contributions is not significant. This suggests that the general population sample’s transferability is just as bad as the student sample’s. I thus, find a similar result as [Goeschl et al. \(2020, p.6\)](#).

**Result 3.** *There is no significant correlation between average contributions in the abstract public goods game and contributions to the real public good of climate change mitigation—for none of the samples.*

## 5. Conclusion

The initial goal of the experiment was to reproduce specific experiments of GMTV in an online setting using a general population sample. 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 reproducible in both meanings: they are *purely* (despite minor differences one can re-analyze the original data and reach the same conclusions) and *scientifically* (one can gather new data drawn from a different population in a different situation and find similar patterns) reproducible. Second, the online experiment proceeded fluently such that dropouts were no concern. Third,

Table 3: Correlations between first-round-dPGG and VCA behavior

	Tobit Regression censoring below 0 and above 100		
	VCA Donation as Share of Earnings		
	(1)	(2)	(3)
First-period contribution in percent	0.07 (0.18)	−0.17 (0.25)	−0.23 (0.25)
Student (1 = yes)		−48.68** (20.44)	−31.06 (21.74)
First-period contr. x Student	4.00*** (0.11)	3.96*** (0.11)	3.94*** (0.11)
Age			0.92** (0.45)
Female (1 = yes)			4.98 (10.74)
Risk Aversion (1-12)		0.46 (0.34)	0.50 (0.34)
Constant	3.16 (10.70)	29.35** (14.67)	−13.93 (26.41)
Observations	116	116	116
Log Likelihood	−359.78	−356.07	−354.01
Akaike Inf. Crit.	725.55	722.15	722.02
Bayesian Inf. Crit.	733.82	735.91	741.29

*Note:*

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01



contribution behavior in the dynamic abstract setting is not linked to behavior in the real world—neither for students nor a more representative sample.

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—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 effort because it did not affect the transferability of abstract results to the real world.

Taken together the results answer the questions whether GMTVs’ dynamic paradigm is reproducible across samples (yes) and whether it is feasible to employ it online (also yes). The question that is *not* answered is, whether GMTVs’ treatment effects replicate well because I did not focus on treatments and treatment effects. It thus, remains an open question, whether the causal effects of GMTVs’ treatments replicate well. Further, we do not know whether their treatments affect real-world behavior. So, how generalizable are treatment effects in public goods games in general? Albeit this study does not address these questions, it provides methods suited to investigate them—even with inconvenient samples.

## 6. Acknowledgements

I gratefully acknowledge support by the German Research Foundation (DFG) under Germany’s Excellence Strategy, cluster EXC 2037 “Climate, Climatic Change, and Society” (project 390683824). I thank Stefan Traub, whoever proofreads this and the participants of the CLICCS B5 group for helpful discussions.

Table 4: 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

## 7. A: Pure Replication

This section comments on two errors as well as a misconception I found in the original data.<sup>20</sup> 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.

### 7.0.1. 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 4 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, participants 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.

### 7.0.2. 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 5 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.

<sup>20</sup>The data can be found in the supplementary materials they provide in their [online appendix](#).

Table 5: 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  $\text{MEAN}=\text{sum}/\text{gdp}$  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.

### 7.0.3. 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 not 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.<sup>21</sup>

---

<sup>21</sup>Accordingly, the definition of **GINI** I provide in Table 4 is not the definition I used to calculate the current period’s Gini coefficient but the previous period’s Gini coefficient.

Table 6: The average amount of tokens contributed over time across samples.

Period	Mean: Original	95%-CI: Original	Mean: Reproduction	95%-CI: Reproduction	p-Value
1	42.739	38.367 - 47.111	40.069	34.406 - 45.732	0.561
2	50.826	42.775 - 58.877	40.138	31.557 - 48.719	0.061
3	59.348	45.916 - 72.780	43.207	30.815 - 55.599	0.067
4	71.565	53.030 - 90.100	46.414	29.930 - 62.898	0.014
5	82.348	55.398 - 109.298	55.931	33.101 - 78.761	0.049
6	87.435	45.735 - 129.135	62.448	31.121 - 93.776	0.191
7	91.652	36.859 - 146.446	68.966	33.626 - 104.305	0.740
8	103.087	30.382 - 175.792	69.759	29.621 - 109.896	0.507
9	98.804	20.125 - 177.483	70.448	22.055 - 118.841	0.747
10	73.609	20.019 - 127.198	70.793	15.116 - 126.470	0.606

## 8. B: Tables

Table 7: The average share of current endowments contributed over time across samples.

Period	Mean: Original	95%-CI: Original	Mean: Reproduction	95%-CI: Reproduction	p-Value
1	0.534	0.480 - 0.589	0.501	0.430 - 0.572	0.561
2	0.487	0.418 - 0.557	0.383	0.309 - 0.457	0.051
3	0.441	0.358 - 0.524	0.325	0.245 - 0.406	0.046
4	0.418	0.338 - 0.497	0.278	0.202 - 0.354	0.009
5	0.377	0.285 - 0.470	0.270	0.190 - 0.349	0.063
6	0.300	0.194 - 0.407	0.234	0.153 - 0.315	0.277
7	0.240	0.143 - 0.337	0.217	0.150 - 0.285	0.934
8	0.214	0.122 - 0.306	0.176	0.111 - 0.240	0.593
9	0.165	0.081 - 0.249	0.149	0.087 - 0.212	0.949
10	0.115	0.049 - 0.181	0.121	0.067 - 0.175	0.450

Table 8: The average wealth over time across samples.

Period	Mean: Original	95%-CI: Original	Mean: Reproduction	95%-CI: Reproduction	p-Value
1	103.348	100.966 - 105.729	101.724	98.806 - 104.642	0.605
2	130.435	124.592 - 136.277	123.517	116.597 - 130.437	0.099
3	161.696	149.708 - 173.683	146.517	133.780 - 159.255	0.061
4	199.522	178.632 - 220.412	171.000	150.455 - 191.545	0.033
5	242.217	208.842 - 275.593	200.586	169.020 - 232.152	0.033
6	288.174	235.294 - 341.054	233.448	186.751 - 280.146	0.047
7	335.478	256.590 - 414.367	269.586	205.759 - 333.413	0.072
8	388.739	275.392 - 502.086	305.897	223.488 - 388.305	0.107
9	439.674	290.308 - 589.040	342.759	239.624 - 445.893	0.124
10	478.087	307.892 - 648.282	379.828	251.997 - 507.658	0.136

Table 9: The average Gini coefficient over time across samples.

Period	Mean: Original	95%-CI: Original	Mean: Reproduction	95%-CI: Reproduction	p-Value
1	0.134	0.111 - 0.157	0.131	0.111 - 0.151	0.818
2	0.150	0.118 - 0.182	0.165	0.139 - 0.192	0.733
3	0.171	0.133 - 0.210	0.186	0.153 - 0.219	0.495
4	0.183	0.137 - 0.230	0.188	0.150 - 0.226	0.768
5	0.192	0.141 - 0.243	0.200	0.159 - 0.241	0.606
6	0.193	0.139 - 0.248	0.194	0.149 - 0.239	0.825
7	0.207	0.153 - 0.260	0.211	0.165 - 0.256	0.905
8	0.214	0.160 - 0.268	0.216	0.170 - 0.261	0.985
9	0.223	0.171 - 0.275	0.220	0.173 - 0.266	0.726
10	0.232	0.179 - 0.285	0.218	0.171 - 0.265	0.606

## References

- Akhtar, A., Ye, H., 2023. Reproducibility and Robustness Replicability of Gsottbauer et al. (2022). I4R Discussion Paper Series 29. The Institute for Replication (I4R). URL: <https://EconPapers.repec.org/RePEc:zbw:i4rdps:29>.
- Amir, O., Rand, D.G., Gal, Y.K., 2012. Economic games on the internet: The effect of \$1 stakes. PLOS ONE 7, 1–4. URL: <https://doi.org/10.1371/journal.pone.0031461>, doi:10.1371/journal.pone.0031461.
- 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/S016726810101950>, doi:[https://doi.org/10.1016/S0167-2681\(01\)00195-0](https://doi.org/10.1016/S0167-2681(01)00195-0).
- Ankel-Peters, J., Fiala, N., Neubauer, F., 2023. Do economists replicate? Journal of Economic Behavior & Organization 212, 219–232. URL: <https://www.sciencedirect.com/science/article/pii/S0167268123001555>, doi:<https://doi.org/10.1016/j.jebo.2023.05.009>.
- Arechar, A.A., Gächter, S., Molleman, L., 2018. Conducting interactive experiments online. Experimental Economics 21, 99–131. URL: <https://doi.org/10.1007/s10683-017-9527-2>, doi:10.1007/s10683-017-9527-2.
- Arechar, A.A., Rand, D.G., 2021. Turking in the time of covid. Behavior Research Methods 53, 2591–2595. URL: <https://doi.org/10.3758/s13428-021-01588-4>, doi:10.3758/s13428-021-01588-4.
- Battaglini, M., Nunnari, S., Palfrey, T.R., 2016. The dynamic free rider problem: A laboratory study. American Economic Journal: Microeconomics 8, 268–308. URL: <https://www.aeaweb.org/articles?id=10.1257/mic.20150126>, doi:10.1257/mic.20150126.
- Bem, D.J., 1987. Writing the empirical journal article, in: The compleat academic. Psychology Press, pp. 171–201. URL: <https://www.taylorfrancis.com/chapters/edit/10.4324/9781315808314-10/writing-empirical-journal-article-daryl-bem>, doi:<https://doi.org/10.4324/9781315808314-10>.
- Benndorf, V., Moellers, C., Normann, H.T., 2017. Experienced vs. inexperienced participants in the lab: do they behave differently? Journal of the Economic Science Association 3, 12–25. URL: <https://doi.org/10.1007/s40881-017-0036-z>, doi:10.1007/s40881-017-0036-z.
- 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>.
- Brick, K., Visser, M., 2015. What is fair? an experimental guide to climate negotiations. *European Economic Review* 74, 79–95. URL: <https://www.sciencedirect.com/science/article/pii/S0014292114001640>, doi:<https://doi.org/10.1016/j.euroecorev.2014.11.010>.
- Brodeur, A., Cook, N., Hartley, J.S., Heyes, A., 2022. Do Pre-registration and Pre-analysis Plans Reduce P-Hacking and Publication Bias? IZA Discussion Papers 15476. Institute of Labor Economics (IZA). URL: <https://ideas.repec.org/p/iza/izadps/dp15476.html>.
- Brodeur, A., Cook, N., Heyes, A., 2020. Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review* 110, 3634–60. URL: <https://www.aeaweb.org/articles?id=10.1257/aer.20190687>, doi:[10.1257/aer.20190687](https://doi.org/10.1257/aer.20190687).
- Brodeur, A., Lé, M., Sangnier, M., Zylberberg, Y., 2016. Star wars: The empirics strike back. *American Economic Journal: Applied Economics* 8, 1–32. URL: <https://www.aeaweb.org/articles?id=10.1257/app.20150044>, doi:[10.1257/app.20150044](https://doi.org/10.1257/app.20150044).
- Buso, I.M., Di Cagno, D., Ferrari, L., Larocca, V., Lorè, L., Marazzi, F., Panaccione, L., Spadoni, L., 2021. Lab-like findings from online experiments. *Journal of the Economic Science Association* 7, 184–193. URL: <https://doi.org/10.1007/s40881-021-00114-8>, doi:[10.1007/s40881-021-00114-8](https://doi.org/10.1007/s40881-021-00114-8).
- Calzolari, G., Casari, M., Ghidoni, R., 2018. Carbon is forever: A climate change experiment on cooperation. *Journal of Environmental Economics and Management* 92, 169–184. URL: <https://www.sciencedirect.com/science/article/pii/S0095069616304557>, doi:<https://doi.org/10.1016/j.jeem.2018.09.002>.
- Camerer, C.F., Dreber, A., Forsell, E., Ho, T.H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H., 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351, 1433–1436. URL: <https://www.science.org/doi/abs/10.1126/science.aaf0918>, doi:[10.1126/science.aaf0918](https://doi.org/10.1126/science.aaf0918), arXiv:<https://www.science.org/doi/pdf/10.1126/science.aaf0918>.
- Carpenter, J., Connolly, C., Myers, C.K., 2008. Altruistic behavior in a representative dictator experiment. *Experimental Economics* 11, 282–298. URL: <https://doi.org/10.1007/s10683-007-9193-x>, doi:[10.1007/s10683-007-9193-x](https://doi.org/10.1007/s10683-007-9193-x).
- 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](https://doi.org/10.1016/j.jbef.2015.12.001).
- Christensen, G., Miguel, E., 2018. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature* 56, 920–80. URL: <https://www.aeaweb.org/articles?id=10.1257/jel.20171350>, doi:[10.1257/jel.20171350](https://doi.org/10.1257/jel.20171350).
- Cook, N.J., Grillos, T., Andersson, K.P., 2019. Gender quotas increase the equality and effectiveness of climate policy interventions. *Nature Climate Change* 9, 330–334. URL: <https://doi.org/10.1038/s41558-019-0438-4>, doi:[10.1038/s41558-019-0438-4](https://doi.org/10.1038/s41558-019-0438-4).
- Eckel, C.C., Grossman, P.J., 1996. Altruism in anonymous dictator games. *Games and Economic Behavior* 16, 181–191. URL: <https://www.sciencedirect.com/science/article/pii/S0899825696900810>, doi:<https://doi.org/10.1006/game.1996.0081>.
- Eichenseer, M., Moser, J., 2019. Leadership in dynamic public good provision: Endogenous growth and inequality. Available at SSRN 3269341 URL: <http://dx.doi.org/10.2139/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](https://doi.org/10.1257/aer.90.4.980).
- Ferraro, P.J., Shukla, P., 2020. Feature—is a replicability crisis on the horizon for environmental and resource economics? *Review of Environmental Economics and Policy* 14, 339–351. URL: <https://doi.org/10.1093/reep/reaa011>, doi:[10.1093/reep/reaa011](https://doi.org/10.1093/reep/reaa011), arXiv:<https://doi.org/10.1093/reep/reaa011>.

- Ferraro, P.J., Vossler, C.A., 2010. The source and significance of confusion in public goods experiments. *The B.E. Journal of Economic Analysis & Policy* 10. URL: <https://doi.org/10.2202/1935-1682.2006>, doi:doi:10.2202/1935-1682.2006.
- Goeschl, T., Kettner, S.E., Lohse, J., Schwioren, 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](https://academic.oup.com/jcr/article-pdf/44/1/196/25496127/ucx047.pdf).
- Gupta, N., Rigotti, L., Wilson, A., 2021. The experimenters’ dilemma: Inferential preferences over populations. [arXiv:2107.05064](https://arxiv.org/abs/2107.05064).
- 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’économie* 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](https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2966.2007.00428.x).
- Hauser, O.P., Rand, D.G., Peysakhovich, A., Nowak, M.A., 2014. Cooperating with the future. *Nature* 511, 220–223. URL: <https://doi.org/10.1038/nature13530>, doi:10.1038/nature13530.
- 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.
- Hänsel, M.C., Bauer, M.D., Drupp, M., Wagner, G., Rudebusch, G., 2022. Climate Policy Curves: Linking Policy Choices to Climate Outcomes. CESifo Working Paper Series 10113. CESifo. URL: [https://EconPapers.repec.org/RePEc:ces:ceswps:\\_10113](https://EconPapers.repec.org/RePEc:ces:ceswps:_10113).
- Inman, M., 2008. Carbon is forever. *Nature Climate Change* 1, 156–158. URL: <https://doi.org/10.1038/climate.2008.122>, doi:10.1038/climate.2008.122.
- Klein, R.A., Ratliff, K.A., Vianello, M., Adams, R.B., Bahník, v., Bernstein, M.J., Bocian, K., Brandt, M.J., Brooks, B., Brumbaugh, C.C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W.E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E.M., Hasselman, F., Hicks, J.A., Hovermale, J.F., Hunt, S.J., Huntsinger, J.R., IJzerman, H., John, M.S., Joy-Gaba, J.A., Barry Kappes, H., Krueger, L.E., Kurtz, J., Levitan, C.A., Mallett, R.K., Morris, W.L., Nelson, A.J., Nier, J.A., Packard, G., Pilati, R., Rutchick, A.M., Schmidt, K., Skorinko, J.L., Smith, R., Steiner, T.G., Storbeck, J., Van Swol, L.M., Thompson, D., van ‘t Veer, A.E., Ann Vaughn, L., Vranka, M., Wichman, A.L., Woodzicka, J.A., Nosek, B.A., 2014. Investigating variation in replicability. *Social Psychology* 45, 142–152. URL: <https://doi.org/10.1027/1864-9335/a000178>, doi:10.1027/1864-9335/a000178, [arXiv:https://doi.org/10.1027/1864-9335/a000178](https://arxiv.org/abs/10.1027/1864-9335/a000178).
- Knuth, D.E., 1984. Literate Programming. *The Computer Journal* 27, 97–111. URL: <https://doi.org/10.1093/comjnl/27.2.97>, doi:10.1093/comjnl/27.2.97, [arXiv:https://academic.oup.com/comjnl/article-pdf/27/2/97/981657/270097.pdf](https://academic.oup.com/comjnl/article-pdf/27/2/97/981657/270097.pdf).
- Krantz, J.H., Ballard, J., Scher, J., 1997. Comparing the results of laboratory and world-wide web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, & Computers* 29, 264–269. URL: <https://doi.org/10.3758/BF03204824>, doi:10.3758/BF03204824.
- Krishna, A., 2021. The need for synergy in academic policies: An introduction to the dialogue on pre-registration. *Journal of Consumer Psychology* 31, 146–150. URL: <https://myscp.onlinelibrary.wiley.com/doi/full/10.1002/jcpy.1211>, doi:doi.org/10.1002/jcpy.1211.
- Milinski, M., Semmann, D., Krambeck, H.J., Marotzke, J., 2006. Stabilizing the earth’s climate is not a losing game: Supporting evidence from public goods experiments. *Proceedings of the National Academy of Sciences* 103,



- 3994–3998. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.0504902103>, doi:10.1073/pnas.0504902103, arXiv:<https://www.pnas.org/doi/pdf/10.1073/pnas.0504902103>.
- Olken, B.A., 2015. Promises and perils of pre-analysis plans. *Journal of Economic Perspectives* 29, 61–80. URL: <https://www.aeaweb.org/articles?id=10.1257/jep.29.3.61>, doi:10.1257/jep.29.3.61.
- Page, L., Noussair, C.N., Slonim, R., 2021. The replication crisis, the rise of new research practices and what it means for experimental economics. *Journal of the Economic Science Association* 7, 210–225. URL: <https://doi.org/10.1007/s40881-021-00107-7>, doi:10.1007/s40881-021-00107-7.
- Paolacci, G., Chandler, J., 2014. Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science* 23, 184–188. URL: <https://doi.org/10.1177/0963721414531598>, doi:10.1177/0963721414531598, arXiv:<https://doi.org/10.1177/0963721414531598>.
- 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.
- Peyton, K., Huber, G.A., Coppock, A., 2022. The generalizability of online experiments conducted during the covid-19 pandemic. *Journal of Experimental Political Science* 9, 379–394. doi:10.1017/XPS.2021.17.
- Pham, M.T., Oh, T.T., 2021. Preregistration is neither sufficient nor necessary for good science. *Journal of Consumer Psychology* 31, 163–176. URL: <https://doi.org/10.1002/jcpy.1209>, doi:10.1002/jcpy.1209.
- Rockenbach, B., Wolff, I., 2017. The effects of punishment in dynamic public-good games. TWI Research Paper Series 106. Thurgauer Wirtschaftsinstitut, Universität Konstanz. URL: <https://EconPapers.repec.org/RePEc:twi:respas:0106>.
- Simmons, J.P., Nelson, L.D., Simonsohn, U., 2021. Pre-registration is a game changer. but, like random assignment, it is neither necessary nor sufficient for credible science. *Journal of Consumer Psychology* 31, 177–180. URL: <https://doi.org/10.1002/jcpy.1207>, doi:10.1002/jcpy.1207.
- Simonsohn, U., Simmons, J., Nelson, L., 2023. Data falsificada (part 1): "clusterfake". URL: <https://web.archive.org/web/20230620144104/https://datacolada.org/109>. accessed on June 20, 2023.
- Snowberg, E., Yariv, L., 2021. Testing the waters: Behavior across participant pools. *American Economic Review* 111, 687–719. URL: <https://www.aeaweb.org/articles?id=10.1257/aer.20181065>, doi:10.1257/aer.20181065.
- Tavoni, A., Dannenberg, A., Kallis, G., Löschel, A., 2011. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences* 108, 11825–11829. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.1102493108>, doi:10.1073/pnas.1102493108, arXiv:<https://www.pnas.org/doi/pdf/10.1073/pnas.1102493108>.
- Vicens, J., Bueno-Guerra, N., Gutiérrez-Roig, M., Gracia-Lázaro, C., Gómez-Gardeñes, J., Perelló, J., Sánchez, A., Moreno, Y., Duch, J., 2018. Resource heterogeneity leads to unjust effort distribution in climate change mitigation. *PLOS ONE* 13, 1–17. URL: <https://doi.org/10.1371/journal.pone.0204369>, doi:10.1371/journal.pone.0204369.
- Waldron, S., Allen, C., 2022. Not all pre-registrations are equal. *Neuropsychopharmacology* , 1–3URL: <https://www.nature.com/articles/s41386-022-01418-x>, doi:<https://doi.org/10.1038/s41386-022-01418-x>.
- Zelmer, J., 2003. Linear public goods experiments: A meta-analysis. *Experimental Economics* 6, 299–310. URL: <https://doi.org/10.1023/A:1026277420119>, doi:10.1023/A:1026277420119.
- Zhou, H., Fishbach, A., 2016. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology* 111, 493–504. URL: <https://doi.org/10.1037/pspa0000056>, doi:10.1037/pspa0000056.