Sustainability as a dynamic game*

Berno Buechel[†, ♯] Corinne Dubois[‡, ♯], Stephanie Fuerer[♯], Tjaša Maillard-Bjedov[♯]

[♯]Department of Economics, University of Fribourg, 1700 Fribourg, Switzerland

April 15, 2025

Abstract

Sustainability is a fundamental concept in the environmental domain, but also in other domains, e.g., regarding personal health. Sustainability means using resources today in a way that does not compromise the availability of resources tomorrow. We propose and test a model that incorporates the essential features of sustainability. First, our Sustainability Game is dynamic in the sense that the actions played in each period have consequences for future periods. Second, there is a contribution threshold that must be reached in order to maintain the level of resources, while some use of resources can be absorbed. Third, it incorporates that the temptation to over-use resources is strong when more than one individual is involved. We first derive equilibrium behavior analytically and then test these pre-registered predictions in the lab. Our main results are the following: (i) Theoretically and experimentally, strategic interaction reduces cooperative behavior and undermines sustainability. (ii) Theoretically and experimentally, lowering the threshold fosters cooperative behavior (i.e., contributing according to the threshold) and sustainability. Our results suggest that technological advancements that lower the threshold for sustainability and behavior change toward sustainability need not be viewed as alternatives, but rather as complementary.

Keywords: Sustainability, conservation, dynamic game, social dilemma, threshold, cooperation

JEL Classification Codes: C73, C92, H41, Q56

*We gratefully acknowledge funding from the University of Fribourg Research Pool (decision of 2020-03-20). We thank Sebastian Berger, Mehdi Farsi, Jana Freundt, Holger Herz, Dominik Karos, Francisco Gomez Martinez, Wojtek Przepiorka, Emanuel Vespa, Christian Zihlmann, and two anonymous referees for their very helpful comments; and Qingchao Zeng for excellent research assistance.

[†]Corresponding author. University of Fribourg, Department of Economics, Bd. de Pérolles 90, CH-1700 Fribourg, Switzerland. Email address: berno.buechel@unifr.ch, web: www.berno.info. Declarations of interest (all authors): none.

[‡]Wüest Partner, 1204 Geneva, Switzerland.

1 Introduction

Sustainability has become the cornerstone of many policy goals, project plans and personal decisions. The United Nations define sustainability as "meeting the needs of the present without compromising the ability of future generations to meet their own needs". In other words, considering that not only future generations but also the same generation in the future can be affected, sustainability means using resources today in a way that does not compromise the availability of resources tomorrow. We develop a model to study sustainability theoretically and experimentally. Our Sustainability Game has three essential features that make it suitable for studying this concept. First, it is dynamic so that decisions made today have an impact on the availability of resources in the future. Second, it includes threshold effects, such that a limited amount of resource utilization can be absorbed and has no impact on the future, but an over-use (i.e., usage above the threshold) leads to a decline in future resources. Third, the game features a tension between private and collective interests when there are multiple decision makers. As discussed extensively below, combining these features sets our model apart from most of the pre-existing literature. The way our model captures sustainability fits particularly well to problems of *conservation*, be it concerning personal health or concerning a natural resource that cannot grow arbitrarily.

We first derive theoretical predictions from this Sustainability Game by studying its equilibria. As an equilibrium concept, we use Markov Perfect Equilibrium (MPE), following, e.g., Vespa (2020).[1] We then test these predictions in a pre-registered laboratory experiment. In the experiment, we focus on varying two dimensions: the number of decision makers and the sustainability threshold. We divide participants into three treatment groups. In the baseline treatment (*T-Baseline*) there are four decision makers, while in the *T-OnePlayer* treatment there is only one. Comparing the two treatments allows an assessment of the extent to which free-riding incentives impact sustainability. In many applications, such as climate and environment preservation, cooperation between many individuals is required to reach a sustainable path, and the temptation to free ride and over-use resources is strong. There are however also instances where an individual is solely involved in maintaining a resource. One such example is personal health. An individual can refrain from unhealthy habits such as smoking, drinking, or eating sugary or fatty products and thereby maintain a good health level. Our model enables us to compare how an individual manages sustainability compared to the case of multiple decision makers involved. The results (both theoretical and experimental) clearly show that players are more likely to reach the sustainability threshold when they are solely responsible for the decision and they are also less likely to contribute zero.

We then analyze how varying the sustainability threshold affects the strategies chosen by players. The treatment *T-LowThreshold* features a lower sustainability threshold than *T-Baseline*, which should theoretically lead to more cooperation (i.e., contributing according to the threshold), as mutual cooperation becomes an equilibrium. Our experimental results confirm the theoretical prediction: a lower threshold increases cooperation and sustainability, while it reduces defection (i.e., contributing zero). In the context of climate change, lowering the threshold could be interpreted as an improvement of carbon capture technology to absorb CO2 emissions, for example. It would lower the effort required by the population to actually meet a sustainability threshold. Whether such a technology undermines (Anderson and Peters, 2016) or fosters (Lackner et al., 2016) emission reduction efforts is an important debate, to which our results offer insights.

Finally, we investigate how specific personal characteristics affect the choice of strategies in the

[1] MPE is a strong equilibrium concept in the sense that it reduces the number of equilibria in dynamic settings. Its main assumption is that strategies do not depend on the history of play other than through the current state variables.

Sustainability Game. We find that agents who score higher on agreeableness tend to contribute more, in particular when contributing zero is the only equilibrium strategy – resonating the findings of, e.g., Volk et al. (2011). Our results also show that agents with higher cognitive ability more often play equilibrium strategies. Specifically, we find for situations with several feasible equilibria, that participants with high cognitive ability are more likely to choose the socially optimal equilibrium – resonating the findings of, e.g., Proto et al. (2019). Finally, when controlling for agreeableness, cognitive ability and further characteristics, cooperation in our Sustainability Game still correlates positively with pro-environmental orientation, yielding suggestive evidence for the external validity of our setup.

This paper's contribution is multi-fold. First, it brings together essential features of sustainability in one tractable model. In particular, our proposed Sustainability Game is dynamic, features a threshold challenge, and incorporates a social dilemma when multiple decision makers are involved. It contributes to the literature (discussed in the next section and illustrated in Table 1) not only by providing a novel social dilemma experiment, but also by providing a new way to elicit sustainable behavior. Second, our first theoretical and experimental result shows that there is a tension between individual free-rider incentives and collective outcomes, which is resolved if there is only one player. The fact that our game can be played meaningfully by one person, hence highlights the core problem that is common to all social dilemma games. This shows that our model "works," not only because it captures the social dilemma aspect, but also because the theoretical predictions from the equilibrium analysis are strongly supported by the experimental results. Third, we address how threshold effects trigger sustainable behavior. We find that a lower threshold increases cooperation and reduces defection.[2] Since defection is still an equilibrium, this result shows that there are participants who target the efficient equilibrium when it comes to equilibrium selection. It also speaks to the crucial debate whether technological advances and behavior change are rather substitutes or complements (see, e.g., Anderson and Peters, 2016, versus Lackner et al., 2016, in the context of CO2 emissions). A major concern of the 'substitutes perspective' is that the benefits of a better technology are undermined by a behavioral response, e.g., that improved energy efficiency might be offset by higher usage (Brockway et al., 2021). On the other hand, the 'complements' perspective emphasizes that a combination of both technological progress and behavior change is necessary to act sustainably (as expressed, e.g., by IPCC experts, https://www.ipcc.ch/2022/04/04/ipcc-ar6-wgiii-pressrelease). Our result, albeit based on a stylized model, yields supportive evidence for the latter view.

The final contribution concerns how behavior in this Sustainability Game is related to personality traits. We predict and show that contributions correlate positively with a participant's agreeableness. Moreover, in situations of equilibrium selection, cooperative behavior correlates positively with cognitive ability. Our results hence suggest three remedies for any challenge of sustainability: first, the problem is relaxed when a single decision maker is made responsible. Second, the problem is relaxed when there is a technology that lowers the threshold sufficiently – and participants are smart enough to choose the efficient equilibrium. Third, the problem is relaxed even if the first two remedies are not available, if the involved decision makers are agreeable people.

The remainder of this paper is organized as follows. In Section 2, we discuss the relation to the literature. In Section 3, we present the theoretical model and characterize its equilibria. In Section 4, we describe the experimental design. In Section 5, we report the experimental results. Section 6 concludes.

[2] In the literature on threshold public goods games, the parallel effects are discussed in particular in Cadsby and Maynes (1999). For threshold public goods games, it is less surprising that higher thresholds reduce cooperation because expected marginal returns are zero up to the threshold and, hence, high thresholds may simply appear out of reach.

2 Related Literature

There are two classic approaches to model social dilemmas: public goods (PG) games and common pool resource (CPR) games. The literature on PG games mostly focuses on a *static* environment to test cooperation and self-interested behavior under different assumptions. The general result is that individuals fail to fully cooperate and under-contribute to the collective good even when interactions are repeated. For reviews of the literature, see Ledyard (1995), Chaudhuri (2011) and Dal Bó and Fréchette (2018). A static environment, be it repeated or not, is however insufficient for analyzing sustainability. Dynamics is essential: the chosen actions today, if unsustainable, affect the possible actions in the future.[3] Moreover, most PG games do not embody a threshold that can be used to distinguish sustainable behavior from unsustainable behavior, the next essential aspect of sustainability. The sub-literature on threshold PG games (e.g., Cadsby and Maynes, 1999; Croson and Marks, 2000) builds a notable exception. These models seem better suited to address sustainability, but most of them are not dynamic either. Moreover, in threshold PG games the expected marginal returns to contributions are zero up to the threshold. We want to model situations where contributions have a positive return, even if the threshold is not reached, because for insufficient contributions to sustainability, the level of these contributions may still matter. An important feature of many threshold PG games and of our model is zero expected marginal returns after the threshold. This models the absorption capacity of health or of a natural resource: different small levels of extraction do not change the resource.[4]

The common pool resource (CPR) games go back to the problem of the commons (see Levhari and Mirman, 1980, Ostrom, 1990, and Walker et al., 2000). Our setup differs from the classic common pool resource (CPR) game in several ways. First and foremost, our game includes threshold effects while the CPR game features a continuous growth, respectively decay, of the pool. This means in particular that a CPR game does not include the possibility that different small levels of usage are absorbed in the sense that they do not affect the size of the resource. Rather, in CPR games the resource grows if there is no extraction and it grows differently for different small levels of extraction. The CPR game is often used for studying renewable resources such as fisheries or forests. Our setup better represents situations in which threshold effects are important. In the case of global warming, for example, experts argue that a certain level of CO2 emissions can be absorbed, while only above a threshold negative consequences occur. Second, in our game, the depletion of resources due to over-use is irreversible. This feature is also well-suited to study environmental issues. In the context of global warming experts predict some adverse consequences of global warming that are irreversible, such as permafrost thaw, the increase in the levels of the oceans or species extinction (see Portner et al., 2022).[5]

More recently, dynamic PG games and variations of CPR games are being considered in the literature. Battaglini et al. (2016) study free-riding incentives in a durable PG game and compare the evolution of the durable public good when investment in it is reversible or irreversible. Gächter et al. (2017) (and similarly Rockenbach and Wolff, 2019) investigate a PG game where the current endowments depend on the past actions. Generally, this dynamic aspect makes cooperation harder to sustain (a classic finding in that respect is provided by Herr et al., 1997). Vespa (2020) analyzes the selection of strategies in a CPR game. He shows that the Markov Perfect Equilibrium (MPE) is the modal strategy in this context. Finally, Przepiorka and Diekmann (2020) study a variation

[3] Dynamics also matters in several applications of PG games that are not related to sustainability (e.g., Cadigan et al., 2011).

[4] The threshold effect for health can be illustrated as follows: A person who needs eight hours of sleep a night cannot improve their health by sleeping thirteen hours, but it makes a difference whether the person sleeps four or five hours.

[5] Irreversible damage and capacity for absorption are also present in the domain of health (e.g., aging is considered as the accumulation of health deficits, Strulik and Grossmann, 2024).

Model	Reference	Dynamic	Recurring	Threshold	Absorption	Irreversible	Intra-generational
Public Goods (PG) Game	e.g., Ledyard (1995), Chaudhuri (2011), Dal Bó and Fréchette (2018)	no	yes	no	no	no	yes
Threshold PG Game	e.g., Cadsby and Maynes (1999), Croson and Marks (2000)	no	yes	yes	yes	no	yes
Common Pool Resource (CPR) Game	e.g., Walker et al. (2000), Vespa (2020), Przepiorka and Diekmann (2020)	yes	yes	no	no	no	yes
Dynamic Free Rider Problem	Battaglini et al. (2016), Gächter et al. (2017), Rockenbach and Wolff (2019)	yes	yes	no	no	both	yes
Collective-Risk Social Dilemma	Milinski et al. (2008)	yes	no	yes	yes	yes	yes
Intergenerational Goods Game	Hauser et al. (2014)	yes	yes	yes	yes	yes	no
Threshold CPR Games	Walker and Gardner (1992), Kimbrough and Vostroknutov (2015)	yes	yes	yes	no	yes	yes
Sustainability Game	This paper	yes	yes	yes	yes	yes	yes

Table 1: Comparison of our modeling approach to the experimental literature.
Notes: Model: Name tag as in original article(s) (if applicable). Reference: original article(s) or survey thereof. Dynamic: State that evolves. Recurring: infinitely recurring challenge (in contrast to finite sequence of actions in one challenge). Threshold: Discontinuity at given threshold of contributions. Absorption: Some extraction is absorbed (e.g., contributions above threshold have the same consequences for the public good). Irreversible: Contributions (e.g., below threshold) can have irreversible consequences. Intra-generational: Externalities concerning the same generation of players (in contrast to externalities concerning future generations of players).

of the CPR games, where a negative externality cumulates over time. Our theoretical framework and laboratory study crucially differs from these contributions in its introduction of a threshold.

Looking for dynamic social dilemma games that incorporate a threshold we found three modeling approaches:[6] First, in the "collective-risk" game introduced by Milinski et al. (2008) and further studied, e.g., by Tavoni et al. (2011) and Szekely et al. (2021), players sequentially decide how much to contribute. If at the end of ten periods the sum of contributions does not reach a given threshold, there is the risk that all payoffs are lost. This model is dynamic, as the contributions accumulate over time, but it is finite, as the challenge is to reach a single threshold in a fixed number of periods. In contrast, our model studies agents who face a similar challenge in every period, while a state variable evolves over time. The second approach for a dynamic model with a threshold is the "intergenerational goods" game introduced by Hauser et al. (2014), where a common resource pool is handed over from one generation to the next. If one generation extracts from the pool more than a given threshold, the pool is not refilled for the next generations of decision makers. A special feature of this game is that it considers the interaction between overlapping generations, while in most other games, players stay the same. Finally, there are also CPR games that do introduce a threshold. The classic model by Walker and Gardner (1992) features even two thresholds. The first one incorporates the idea of a "safe yield" zone, a (small enough) level of extraction where the probability of destruction of the resource is zero. Similar in spirit but different in consequences, we consider a safe zone, where the resource maintains its size for different (small enough) levels of extraction. This is also a difference to the threshold CPR model provided by Kimbrough and Vostroknutov (2015). In their model, the returns to extracting the resource depend on whether the collective level of extraction is below or above the given threshold, while the resource grows for small amounts of extraction. Hence, both models differ from ours by capturing a different view on absorption capacity. That is why our model fits best to problems of conservation and less to problems of a resource that grows when there is no extraction.

Table 1 compares our model to modeling approaches from the experimental literature. As the table shows, having a model that is dynamic and has a threshold, already distinguishes our model from several approaches (see the four rows starting with PG games down to the dynamic free rider problem). The three approaches that share these features (the collective-risk social dilemma, the intergenerational goods game, and the threshold CPR games) differ in other respects as discussed immediately above. Our game contributes to the literature by incorporating important aspects of many sustainability problems, as highlighted in Table 1. Despite these features, our model is simple enough to be solved analytically and to be implemented in laboratory settings, as we demonstrate.

Concerning the findings, our results are in line with experimental studies showing that cooperation is more likely when it is an equilibrium and especially when it is the unique equilibrium of the game.[7] More specifically, they are in line with threshold PG games, in which lower thresholds compared to the rewards of reaching them increase relative contributions (Cadsby and Maynes, 1999; Croson and Marks, 2000). Our results show that this also holds in the Sustainability Game; but the difference to the treatment *T-Baseline*, where defection is the unique equilibrium, becomes smaller with learning. Our model also adds to a literature that studies how personal traits affect cooperative behavior. Cognitive ability is related to playing close to equilibrium and to efficient equilibrium selection (Gill and Prowse, 2016; Proto et al., 2019, 2022), while agreeableness is as-

[6] Lange (2022, Table 5) provides an overview of social dilemma experiments that are used to measure pro-environmental behavior. His table shows that many of the experimentally studied games are either dynamic or incorporate a threshold. It also reveals that those who have both features, relate back to the modeling approaches of Milinski et al. (2008) and Hauser et al. (2014) that we discuss. Lange (2022)'s survey also includes various measures of pro-environmental behavior with real-world consequences. Fixing such a consequence helps to make the external validity credible, but clearly restricts the domain of applications.

[7] Moreover, these games are more likely to be chosen when participants are allowed to choose between games (e.g., Dannenberg et al., 2011).

sociated with cooperative behavior and with positive reciprocity (e.g. Volk et al., 2011; Dohmen et al., 2008).

3 The Model

Our Sustainability Game features n agents who interact for an infinite number of periods. At the beginning of every period $t = 0, 1, 2, ...$, each player $i = 1, 2, ..., n$ receives an endowment e_t. Each player i must then decide what share $c_{i,t}$ thereof she contributes to a special account; the remaining part of the endowment, $(1 - c_{i,t})e_t$, goes to her private account. After each period t, the game continues to period $t + 1$ with probability $\delta \in (0, 1)$, and ends with probability $(1 - \delta)$.

If the game continues to period $t + 1$, the total amount put on the special account by all n players in period t determines the endowment in period $t + 1$ in the following way:

$$e_{t+1} = \begin{cases} e_t & \text{if } \sum_{i=1}^{n} c_{i,t} e_t \geq Z_t \\ e_t - g\left(Z_t - \sum_{i=1}^{n} c_{i,t} e_t\right) & \text{else,} \end{cases} \tag{1}$$

where $Z_t = zne_t$ is the sustainability threshold set by a parameter $z \in [0, 1]$, and $g \in (0, \frac{1}{zn}]$ is a loss parameter. In words, the sustainability threshold parameter z defines which fraction of the overall endowment needs to be contributed to the special account in order to maintain the level of endowment from one period to the next. If the group as a whole contributes the sustainability threshold or more, the endowment in the next period will be identical to the current endowment. If the sustainability threshold is not met, the next endowment will decline proportionately to the shortfall, according to loss parameter g. Linearity of decline is assumed since it seems to be the simplest functional form. Note that the initial endowment e_0 is exogenous but endowments in all subsequent periods depend on players' decisions in past periods. Each player's payoff consists of the amounts kept in her private account.

Two specific characteristics of the game are worth mentioning. First, since the sustainability threshold is proportional to endowment, the situation is the same in relative terms for any endowment level and at any time.[8] Second, the reduction in endowment, if it happens, is irreversible. A special case occurs when $g \equiv \frac{1}{zn}$, as contributing zero by all players then fully depletes the next endowments. While our theoretical analysis is not restricted to this special case, we will focus on it in the laboratory experiment.

3.1 The social planner's solution

Imagine a benevolent social planner could decide which fraction c_t of the aggregate endowment ne_t the whole group contributes to the special account in every period t. This social planner maximizes the aggregate expected payoff $EP_0 = \sum_{t=0}^{\infty} \delta^t (ne_t - c_t ne_t)$ over the contribution share c_t. The solution to the social planner's problem delivers the socially optimal allocation of the Sustainability Game.

Proposition 1 (Social Planner's Solution)**.** *Suppose $\delta > \underline{\delta} := \frac{1}{1+ng(1-z)}$. The socially optimal contribution is $c_t^* = z$ in every period t.*

This proposition is proven in Appendix A.1. It means that if the probability of reaching the following period is not too small, contributing the threshold – and thereby investing in future endowments without wasting contributions – maximizes social welfare. This solution determines the optimal average contribution suggesting, but not requesting, that every player contributes

[8] While realism of this assumption depends on the application, it certainly buys tractability and parsimony.

equally. We are interested in problems where it is worthwhile from a social perspective to behave in a sustainable manner and therefore mostly restrict our analysis to cases where $\delta > \underline{\delta}$.

3.2 Markov perfect equilibria

To study individually rational contributions in the Sustainability Game described above, we look at symmetric Markov-perfect equilibria (MPE). MPE are a subset of sub-game perfect equilibria in which agents use stationary Markov strategies. Markov strategies do not depend on past decisions taken in a game, other than through the current levels of the state variables. In our game, the only state variable is the endowment. Each player thus maximizes her expected payoff conditional on the state variable e_t, which evolves according to Equation 1, and taking the actions of other players as given:

$$\max_{c_{i,t}} EP_{i,0} = \sum_{t=0}^{\infty} \delta^t \left(e_t - c_{i,t} e_t\right). \tag{2}$$

Proposition 2 (Symmetric Markov Perfect Equilibria)**.**

1. *If* $\delta < \bar{\delta} := \frac{1}{1+g(1-z)}$*, contributing zero (*Defection*) is the unique symmetric MPE.*
2. *If* $\bar{\delta} \leq \delta \leq \bar{\bar{\delta}} := \frac{1}{1+g(1-zn)}$*, there are three symmetric MPE (for* $n > 1$*):*
   - *i. contributing* $c_{i,t} = 0$ *(*Defection*),*
   - *ii. contributing* $c_{i,t} = z$ *(*Cooperation*),*
   - *iii. contributing* $c_{i,t} = \frac{1-\delta(1-g(zn-1))}{\delta g(n-1)} \in [0, z]$ *(*Inbetween*).*
3. *If* $\delta > \bar{\bar{\delta}}$*, contributing* $c_{i,t} = z$ *(*Cooperation*) is the unique symmetric MPE.*

The proposition is proven in Appendix A.2. It shows that for low discount factor δ (while still $\delta > \underline{\delta}$) the Sustainability Game is a social dilemma situation, in which individual incentives are to contribute zero (*Defection*), while the collective optimum is achieved with contribution of z (*Cooperation*) by Proposition 1. For a high discount factor, *Cooperation* is the unique equilibrium and for an intermediate discount factor both behaviors are symmetric MPE, while there is an additional equilibrium with intermediate contributions. This additional equilibrium, *Inbetween*, is weak in the sense that individual deviations do not reduce utility of the deviating agent, but leave her indifferent (as shown at the end of the Proof of Proposition 2). For the special case that there is only a single player ($n = 1$), we have $\underline{\delta} = \bar{\delta} = \bar{\bar{\delta}}$. Then the individually and socially optimal behavior is *Cooperation* (*Defection*) for a discount factor above (below) these thresholds.

Note that in the Sustainability Game, players' contributions are typically strategic complements as long as the aggregate contribution lies below the threshold. As soon as the threshold is reached, the contributions of different players become strategic substitutes.

4 Experimental Design

This pre-registered experiment consists of two parts. In the first part, participants play the Sustainability Game six times. We use a between-subject design where each participant is assigned to one out of the three treatments called *T-Baseline*, *T-OnePlayer*, and *T-LowThreshold*. The treatments differ from one another with respect to (i) the number of players and (ii) the sustainability threshold, as explained below. The second part of the experiment is identical for all treatments and used to elicit personal characteristics of the players, including cognitive ability, agreeableness, risk aversion, ecological attitudes and some socio-demographic characteristics.

4.1 Implementation and procedures

We programmed the experiment using the o-Tree framework of Chen et al. (2016), and ran the sessions online on the Prolific platform between January 19$^{\text{th}}$ and January 25$^{\text{th}}$ 2022. We ran a total of 18 sessions (6 per treatment), with a minimum of 4 and a maximum of 7 groups or single players per session. The assignment of sessions to treatments was randomly drawn by the computer. Participants were provided with detailed instructions for each part of the experiment and we ensured their understanding of the Sustainability Game with a comprehension questionnaire.[9] The study duration was around 60 minutes. A total of 282 participants completed the study and earned on average 7.80 GBP.[10]

For each participant the timeline is as follows. First she reads the instructions and completes the comprehension questionnaire. Then she is matched into a group of $n = 4$ or stays single ($n = 1$), depending on the treatment. Then she plays the Sustainability Game for the first time. In each period of the game she decides how to allocate her current endowment between the private and the special account, i.e., how much to contribute. The game continues from one period to the next with probability $\delta = 0.65$; which yields $\frac{1}{0.35} \approx 2.86$ periods in expectation. Technically, the game is played in blocks of five periods and when the end of the game is drawn within a block, the participant is informed after the block. This means that each participant certainly makes decisions in the first five periods.[11] After the end of the first game, the participant is informed about her payoff of this game. Now, the second game starts. After six games with the same group composition, this part ends. In the second part the participant completes a survey about personal characteristics. Finally, one game is randomly selected to be payoff relevant and the payoff is received.

4.2 Social dilemma treatment *T-Baseline*

Participants are matched into groups of four players ($n = 4$). Each player receives an initial allocation of 100 points. Then each player decides how to allocate this endowment between the private and the special account, i.e., how much to contribute to the special account. At the end of the period, she receives information about how much the other members of her group have contributed, the aggregate contribution, whether the threshold was met and how large the next period's endowment is. The game continues to the next period with probability $\delta = 0.65$. Each player faces the new endowment, which either equals the last period's endowment (if the total amount in the special account reached the threshold), or is smaller (if the threshold was not reached), and decides again how to allocate it between the special and the private account. In the whole experiment, the loss parameter is $g \equiv \frac{1}{zn}$, which ensures that zero contributions by all players reduce the next endowments to zero. In the baseline Treatment *T-Baseline*, the sustainability threshold parameter z is set to 0.5. With this parametrization, the endowment evolution of Equation 1 simplifies to $e_{t+1} = \min\{e_t, C_t/2\}$, where C_t is the aggregate contribution of the group in period t.

The parametrization in *T-Baseline* yields $\underline{\delta} = 0.50$ and $\bar{\delta} = 0.80$, which implies $\underline{\delta} < \delta < \bar{\delta}$. By Propositions 1 and 2.1 (where 2.1 stands for the first part of Proposition 2), *T-Baseline* therefore represents a pure social dilemma: it is socially optimal for the group to contribute the threshold

[9] The instructions, the questionnaire and all materials concerning the study design including the pre-registered hypotheses can be found at `https://doi.org/10.1257/rct.6132`. All information had been provided and uploaded until January 18th, 2022 and there were no changes since then.

[10] We only consider groups that played the complete game. Details about attrition are provided in Appendix B.1.

[11] If the end of the game was drawn within this block, say after the third period, then the decisions in period four and five are not payoff-relevant. If the end of the game was not drawn within this block, then this group plays a block of five further periods. This approach, called block random design, follows Fréchette and Yuksel (2017).

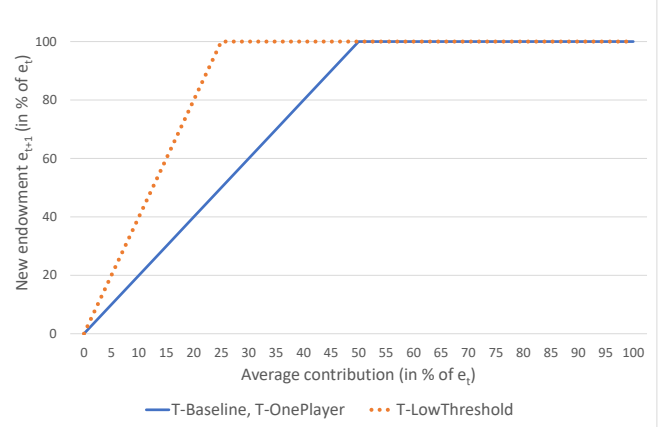
amount but it is individually rational for each player to defect, i.e., contribute zero (independently of the other players' behavior).

4.3 No strategic interactions treatment *T-OnePlayer*

T-OnePlayer eliminates strategic interactions by setting the group size to $n = 1$, but it imitates *T-Baseline* in the other respects. Hence, the sustainability threshold parameter is kept at $z = 0.5$ and zero contributions still lead to full exhaustion. The formula for the evolution of endowments now reduces to $e_{t+1} = \min\{e_t, 2C_t\}$. Figure 1 illustrates the relationship between average contribution and future endowment, which is identical for *T-OnePlayer* and *T-Baseline*. Like all treatments, this treatment has an initial allocation of 100 points and the continuation probability is $\delta = 0.65$.

Under *T-OnePlayer* we have $n = 1$ which yields $\underline{\delta} = \bar{\delta} = \bar{\bar{\delta}} = 0.5 < \delta$. By Propositions 1 and 2.3, *T-OnePlayer* therefore features a unique equilibrium that coincides with the social optimum: *Cooperation* (i.e., contributing according to the threshold).

Figure 1: Next endowment as a function of current average contribution



Notes: This figure shows how next period individual endowment e_{t+1} depends on the average contribution to the special account by the n group members in period t. Both variables are expressed as percent of current individual endowment e_t. This graph applies to any period t because, in relative terms, the stage game is the same in every period.

4.4 Low-threshold treatment *T-LowThreshold*

T-LowThreshold keeps group size $n = 4$ of the baseline, but lowers the threshold. Specifically, the sustainability threshold parameter is set to $z = 0.25$, while zero contributions still lead to full exhaustion. Figure 1 shows that in *T-LowThreshold* it is sufficient to reach an average contribution share of 25% of the endowment to maintain future endowments at the current level, whereas in *T-Baseline* (and in *T-OnePlayer*) an average contribution share of 50% is necessary. The formal expression for the evolution of endowments, Equation 1, now reduces to $e_{t+1} = \min\{e_t, C_t\}$.

The parametrization of *T-LowThreshold* yields $\underline{\delta} = 0.25$, $\bar{\delta} \approx 0.57$ and $\bar{\bar{\delta}} = 1$, which implies $\underline{\delta} < \bar{\delta} < \delta = 0.65 < \bar{\bar{\delta}}$. By Propositions 1 and 2.2, *T-LowThreshold* therefore represents a problem of equilibrium selection, where both *Defection* and *Cooperation* are MPE, while *Cooperation* is still the social optimum. Table 2 summarizes the three treatments and their respective implications for

the equilibria of the game.[12] We considered *T-Baseline* as the natural baseline because it captures a pure social dilemma.

Table 2: Treatments summary

	T-Baseline	*T-OnePlayer*	*T-LowThreshold*
group size n	4	1	4
sutainability threshold param. z	0.5	0.5	0.25
discount factor δ	0.65	0.65	0.65
social optimum	*Cooperation*	*Cooperation*	*Cooperation*
equilibria	*Defection*	*Cooperation*	*Cooperation*, *Defection*, *Inbetween*

Notes: *Cooperation* is defined as contributing according to the threshold, $c_{i,t} = z$. *Defection* is defined as contributing zero, $c_{i,t} = 0$. We restrict attention to symmetric Markov-perfect equilibria.

4.5 Personal characteristics

In the second part of the experiment, we elicit information about various characteristics of the participants. First, participants receive 30 points and must decide how many to invest in a profitable but risky project. We measure risk tolerance with the amount invested. Then, we measure cognitive abilities of participants using 12 questions out of the Set 2 of the Raven Advanced Progressive Matrices. The number of questions a participants answers correctly gives her/his *Raven* score. Participants then take a personality test that consists of 24 items from the International Personality Item Pool, based on Maples-Keller et al. (2019). We measure two aspects of participants' personality: agreeableness and conscientiousness. Agreeableness measures a person's altruism, trust in others, cooperation and morality. Conscientiousness measures self-discipline, efficiency, achievement-striving and dutifulness. We then use the New Ecological Paradigm (NEP) of Dunlap et al. (2000) to assess pro-environmental orientation of participants.[13] After that, participants are asked to complete a short CO2 footprint questionnaire that consists of six questions from the WWF Swiss footprint calculator, following Berger and Wyss (2021). Our study finishes with a standard short demographic questionnaire.

4.6 Hypotheses

Before running the experiments, we had pre-registered four primary hypotheses and two secondary hypotheses.[14] The primary hypotheses concern treatment effects (derived from the game-theoretic equilibrium analysis); the secondary hypotheses concern expected correlations between individual traits and behavior.

Let us begin with the primary hypotheses about treatment *T-OnePlayer*. In the baseline setting, *T-Baseline*, *Defection* (i.e., contributing zero) is the unique equilibrium. In contrast, in the setting without strategic interaction, *T-OnePlayer*, *Cooperation* (i.e., contributing z) is the unique equilibrium. Hence, we predict:

Hypothesis 1 (H-CoopDef-One)**.** *Without strategic interaction (i.e., in* T-OnePlayer*),* Cooperation *is more often played and* Defection *is less often played (than in the baseline,* T-Baseline*).*

[12] A combination of one player and low threshold is also thinkable, but it has the same theoretical predictions as the one player treatment.

[13] The New Ecological Paradigm Scale is a revised and extended version of the original New Environmental Paradigm Scale, also abbreviated with NEP.

[14] These hypotheses were pre-registered at AEAregistry-6132. They are copied and pasted here with identical wording and order. We here only add the treatment names in brackets, replacing longer explanations between the hypotheses in the pre-registered 'Analysis plan' document. As Brodeur et al. (2024) document, pre-registration, as practiced by economists, varies in terms of having or not having a pre-analysis plan and its level of specificity. Our analysis plan's specificity can be seen from the following example. To test Hypothesis H-CoopDef-One, we wrote: "Compare frequency of contributing near Cooperation and near zero between T-Baseline and T-OnePlayer."

This game-theoretic prediction captures free-riding incentives that arise when there are multiple players. As usual, free-rider incentives are accompanied by strategic uncertainty and by the aversion of being exploited by free-riders. Since *Cooperation* means reaching the threshold, while *Defection* does not, the behavior predicted in Hypothesis H-CoopDef-One has the following direct consequence:

Hypothesis 2 (H-Reached-One)**.** *Without strategic interaction (i.e., in* T-OnePlayer*), the threshold is reached more often (than in the baseline,* T-Baseline*).*

We now turn to treatment *T-LowThreshold*, which differs from the baseline by its lower threshold. In *T-LowThreshold* both *Defection* and *Cooperation* are equilibria. Thus, participants in that treatment might more often play *Cooperation* than in the baseline *T-Baseline*, where *Defection* is the unique equilibrium. Hence, we hypothesize:

Hypothesis 3 (H-CoopDef-Low)**.** *When the threshold is lower (i.e., in* T-LowThreshold*),* Cooperation *is more often played and* Defection *is less often played (than in the baseline,* T-Baseline*).*

This hypothesized behavior on the individual level has the following consequences on the collective level:

Hypothesis 4 (H-Reached-Low)**.** *When the threshold is lower (i.e., in* T-LowThreshold*), it is reached more often (than in the baseline,* T-Baseline*).*

Notice that Hypothesis H-Reached-Low is in some sense less challenging than the others, as it is additionally supported by a mechanical effect: the same contributions that are insufficient for a high threshold, may be sufficient to reach a small threshold.

In addition to these four primary hypotheses we also pre-registered two secondary hypotheses. First, the standard index of *Agreeableness* that we use includes measures of altruism, cooperation, trust and morality. These aspects of personality should translate into more pro-social behavior in our game and more specifically correlate positively with contributions. Hence, we hypothesize:

Hypothesis 5 (H-Agree)**.** *Agreeable people contribute more.*

Second, people with high cognitive ability might play more often the game-theoretic equilibrium strategies than people with lower cognitive ability. The idea is that high-cognition individuals may think more more strategically, using deeper reasoning about their own and others' incentives. However, equilibrium behavior depends on the behavior of all players involved. If a high-cognition subject expects other group members not to have high cognition, then the best response need not be an equilibrium strategy. Still, we expect participants with high *Raven* score to play more often *Cooperation* (in *T-OnePlayer* and *T-LowThreshold*) and to play more often *Defection* (in *T-Baseline* and *T-LowThreshold*) than the agents with a low Raven score. This is summarized by Hypothesis H-Raven.

Hypothesis 6 (H-Raven)**.** *People with high cognitive ability play more often* Cooperation *and play more often* Defection*.*

5 Experimental Results

Before testing the six hypotheses, we briefly describe the data set.

5.1 Descriptive statistics

We have 282 participants in our study. They are aged between 19 and 63, with an average of around 27 years. 52% of participants declare themselves as women and 48% as men or other (with 1.4% choosing other). On average our participants have a *Raven* score of 6.5 out of 12 with substantial variation between participants, and an agreeableness index of 45.0 (on a scale from 12 to 60). These and further descriptive statistics are summarized in the first block of Table 3.

The participants form 63 groups of four (of which 31 are in treatment *T-Baseline* and 32 in treatment *T-LowThreshold*) and 30 remain single players (those in *T-OnePlayer*). Table 3 summarizes all important outcome variables, first those on the individual level, then those on the collective level. Each game lasts at least five periods and is repeated six times, yielding 2,790 observations on the group level and 8,460 observations on individual behavior.[15] The binary variables *Close to Cooperation*, *Close to Defection*, and *Close to Inbetween*, equal 1 if a participant has submitted a contribution share within a $\pm 2pp$ band around the strategy.[16] Pooling all individual decisions, participants play *Close to Cooperation* in 30.1% of all cases, while they play *Close to Defection* in 7.8% of all cases. The strategy *Close to Inbetween* can only be played in *T-LowThreshold* and is chosen in only 3.2% of these cases. Considering the whole distribution of contribution shares shows that there are no other frequently played strategies and that contributing according to the threshold (*Close to Cooperation*) is by far the modal choice in each treatment (Appendix B.2). *Threshold met* is a binary variable that equals 1 if the participant's group has reached the sustainability threshold in a given period t and zero otherwise. On average, participants reach the sustainability threshold 55% of the time. *Sustainability* is a binary variable that equals 1 if the group reached the threshold in all periods until t, or equivalently, maintained their endowment at its original level of 100.

Figure 2 illustrates the evolution of the main variables over time separated by treatment. In particular, it displays the percentage of participants who played *Close to Cooperation* (upper left panel) and *Close to Defection* (upper right panel), as well as the share of *Threshold met* (lower panel). If anything, there is a downwards tendency in *Close to Cooperation* and an upwards tendency in *Close to Defection*, while *Threshold met* does not seem to systematically change over the periods. Figure 2 also yields a first impression of potential treatment effects.

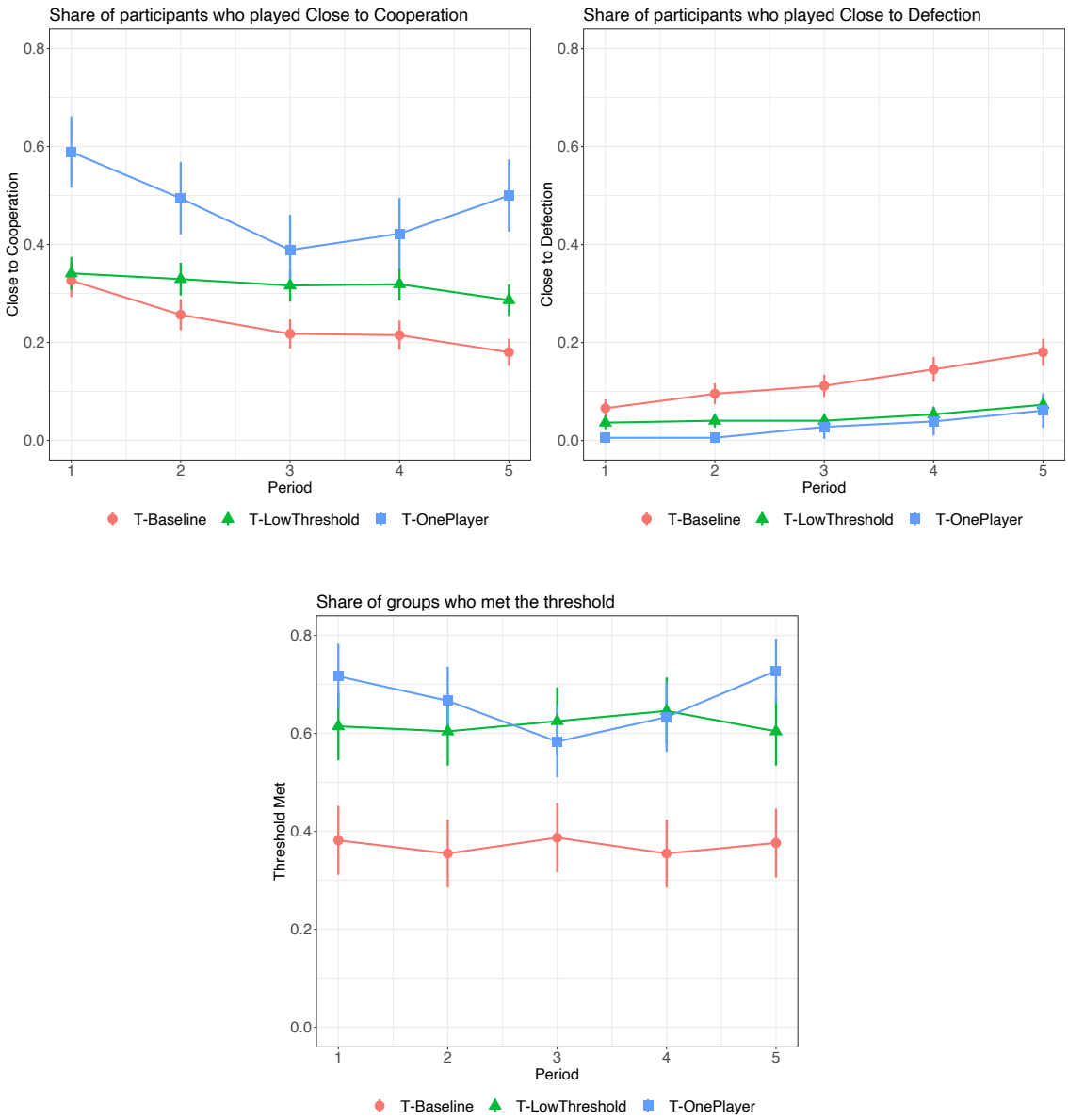
To causally identify treatment effects, we again compare these outcomes in the three treatments but also account for interdependencies of observations. First, the behavior of a group across the five periods of the game is clearly interdependent. Therefore, each analysis uses either only a single outcome period or the (single) average of all five outcome periods. Second, apart from period 1 in the first game, each group has a common history, which can make their choices interdependent. We deal with this dependency by clustering standard errors on the group level, which generally helps when there is intra-group correlation of residuals (see Abadie et al., 2017). Moreover, as a robustness test, we additionally run the main analyses restricting the sample to the first game for each group and explicitly report the first round outcome.[17] Finally, to assure that different behavior across treatments (if any) is not driven by an unlucky draw of allocating different types to treatments or by attrition, we include control variables.

[15] We focus on the first five periods in each game because these can be observed independently of the actual end of the game, thanks to the block random design. Out of 558 games, 481 ended in the first block (periods 1-5), 69 ended in the second block (periods 1-10), and 8 ended in the third block (periods 11-15).

[16] For *T-Baseline* and *T-OnePlayer Close to Cooperation* therefore equals 1 for contribution shares between 0.48 and 0.52 and 0 otherwise. For *T-LowThreshold Close to Cooperation* equals 1 for contribution share between 0.23 and 0.27 and 0 else. *Close to Defection* equals 1 if a participant chose a contribution share between 0 and 0.02. This approach is in the tradition of, e.g., Walker et al. (2000).

[17] In Section 5.5, we explore potential learning effects over repetitions of the game.

Figure 2: Main variables evolution over time, by treatment



Notes: Mean and standard 95% confidence intervals, pooling groups and repetitions of the game. *Close to Cooperation* and *Close to Defection* are binary variables that equal 1 if a participant has submitted a contribution share within a $\pm 2pp$ band around the *Cooperation* and *Defection* strategies, respectively. *Threshold met* is a binary variable that equals 1 if a participant's group has reached the sustainability threshold.

Table 3: Summary statistics

Variable	N	Mean	St. Dev.	Min	Max
Female	282	0.521	0.500	0	1
Age	282	27.372	7.784	19	63
Raven	282	6.518	3.113	0	12
Risk Tolerance	282	14.372	7.403	0	30
Agreeableness	282	44.993	6.380	26	58
Conscientiousness	282	43.323	7.012	24	60
Contribution	8,460	29.549	19.747	0	100
Contribution Share	8,460	0.354	0.217	0.000	1.000
Private account	8,460	51.982	23.724	0	100
Close to Cooperation	8,460	0.301	0.459	0	1
Close to Defection	8,460	0.078	0.268	0	1
Close to Inbetween	3,840	0.032	0.177	0	1
Endowment	2,790	81.904	27.362	0	100
Threshold met	2,790	0.551	0.497	0	1
Sustainability	2,790	0.349	0.477	0	1

Notes: The variable *Raven* measures cognitive ability of the participants and corresponds to the number of questions that they answered correctly in the *Raven* test that consisted of 12 questions. *Risk Tolerance* is the number of points out of 30 invested in the profitable risky project in the risk aversion game, where a higher score indicates higher tolerance to risk. *Agreeableness* and *Conscientiousness* report participants' scores on the International Personality Item Pool test. Observations on individual behavior and on group outcomes are here pooled over all three treatments, all five periods, and all six repetitions of the game.

5.2 Test of primary hypotheses: treatment effects

This first part of our econometric analysis estimates treatment effects on the three main outcome variables, *Close to Cooperation* (Table 4), *Close to Defection* (Table 5), and *Threshold met* (Table 6). We use dummy variables for the treatments *T-LowThreshold* and *T-OnePlayer* and keep *T-Baseline* as the reference category. We run each regression once without and once with controls. The set of control variables is *Age*, *Female*, *Raven*, *Risk Tolerance*, *Agreeableness* and *Conscientiousness*. All regression models report robust standard errors clustered at the group level.

Table 4 reports treatment effects on a participant's probability of playing *Close to Cooperation*. Columns (1) and (2) only consider the first period of each game, Columns (3) and (4) only consider period five. (Recall that for all participants we have observations of their first five periods.) The dependent variable is binary, so we estimate the coefficients using a logistic regression model. The raw coefficients of the logit regressions are the log-odds (which are easily transformable into marginal effects). In Columns (5) and (6), we take the individual's average of the binary variable *Close to Cooperation* over periods one to five of each game. The dependent variable therefore represents the frequency of playing *Close to Cooperation* over the five periods of the game; it is no longer binary but can take values in $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. To easily interpret coefficients, we treat this dependent variable as continuous and run an OLS regression.

The results in Table 4 show that removing strategic interactions significantly increases the probability of playing *Close to Cooperation*. This effect is illustrated in the upper left panel of Figure 3, which shows the regression coefficients when running it for each of the five periods. We observe that the effect of removing strategic interaction occurs already in period one, is stable and persists until period five. Columns (5) and (6) of Table 4 indicate that removing strategic interactions increases the frequency of playing *Close to Cooperation* by about 24-25pp. Transforming the coefficients of the logit regressions into marginal effects yields a similar quantification. For

instance, the coefficient of *T-OnePlayer* in Column (1), 1.083, turns into a marginal effect of 0.262, again indicating an increase of around 25pp (or even larger when using the other coefficients).[18] Reversely formulated, introducing strategic interaction reduces the share of subjects playing close to the efficient strategy by about 25pp, which means that the share is roughly cut in half. The effect is large and statistically significant. Note that the inclusion of the controls does not change the size or significance of the estimates. Finally, we challenge the robustness of these results by re-running the analysis of Table 4, now restricted to the sub-sample of the first game of each group (Appendix Table B.1). The findings are confirmed even with this smaller number of observations. These results fully support Hypothesis H-CoopDef-One, first with respect to *Cooperation*.

The results in Table 4 also indicate that, in line with Hypothesis H-CoopDef-Low, lowering the threshold significantly increases the probability of *Close to Cooperation*, in period five, but not in period one (compared to *T-Baseline*). The left panel of Figure 3 confirms that treatment effects of *T-LowThreshold* increase over time and only become significant toward the end of the first five periods. Columns (5) and (6) of Table 4 suggest that lowering the threshold results in an increase of about 8pp in the frequency of playing *Close to Cooperation*. The weak significance of the *T-LowThreshold* coefficients in Columns (5) and (6) is due to the insignificance of treatment effects in early periods of the game. The robustness test in Table B.1 supports significance of these effects. Hence, Hypotheses H-CoopDef-One and H-CoopDef-Low are largely supported concerning *Cooperation*.

Table 4: Treatment effects on playing Close to Cooperation

Dependent Variable:	Close to Cooperation					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
T-OnePlayer	1.083***	1.275***	1.516***	1.657***	0.2396***	0.2508***
	(0.3396)	(0.3407)	(0.3409)	(0.3518)	(0.0672)	(0.0617)
T-LowThreshold	0.0654	0.0814	0.6030**	0.6451**	0.0792	0.0821*
	(0.2575)	(0.2566)	(0.2849)	(0.2747)	(0.0510)	(0.0461)
Constant	-0.7235***	-1.676	-1.516***	-1.432*	0.2392***	0.1412
	(0.1889)	(1.045)	(0.2285)	(0.8541)	(0.0360)	(0.1545)
Controls	No	Yes	No	Yes	No	Yes
Fit statistics						
Observations	1,692	1,692	1,692	1,692	1,692	1,692
R^2					0.04257	0.13610
Pseudo R^2	0.01973	0.09267	0.03928	0.07353		
Wald (joint nullity)	5.7419	5.3471	9.9380	4.8813	6.3765	7.1612

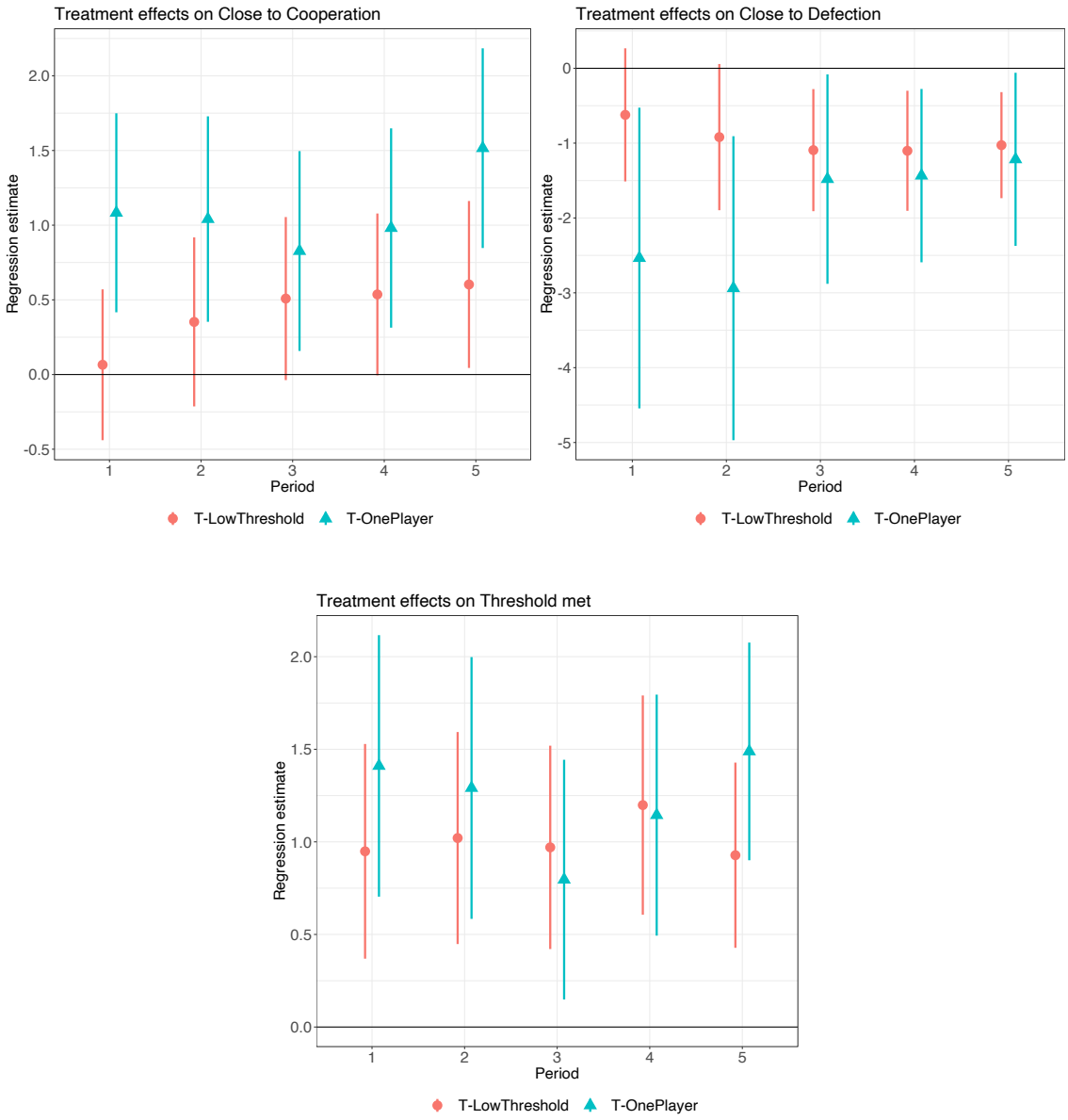
Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Concerning *Defection*, the right panel of Figure 3 illustrates the treatment effects on a participant's probability of playing *Close to Defection*. Table 5 shows the corresponding regressions. Lowering the threshold significantly reduces the probability of *Close to Defection*, again with stronger effect in period five than in period one. The average frequency of playing *Close to Defection* falls by about 8 to 9pp in *T-OnePlayer* and by 7pp in *T-LowThreshold* relative to *T-Baseline*.

[18] We refrain from reporting all marginal effects in the paper, as we have a good assessment of the effect size with Figure 2 that illustrates the differences in outcomes between treatments and with the OLS accompanying all logit regressions (e.g., Columns 5-6 of Table 4).

The results in Table 5 and Figure 3 show that Hypotheses H-CoopDef-One and H-CoopDef-Low are not only supported concerning *Cooperation* but also concerning *Defection*.[19]

Figure 3: Treatment effects relative to *T-Baseline* on strategies played, over time



Notes: Coefficient and 95% confidence intervals from logistic regression with dependent variable *Close to Cooperation* (upper left panel) and *Close to Defection* (upper right panel) and *Threshold met* (lower panel), on *T-LowThreshold* and *T-OnePlayer* dummies, and on a constant. Estimation as in Tables 4-6 with standard errors clustered on the group level.

The remaining primary hypotheses are Hypotheses H-Reached-One and H-Reached-Low, which predict that without strategic interaction, or with a lower threshold, the threshold is reached more often. Table 6 presents the results.[20] It shows that both *T-OnePlayer* and *T-LowThreshold* significantly increase the groups' probability of reaching the threshold in period one and period

[19]Notice that for the strategy *Inbetween* there are no treatment effects to consider, as it is only playable in one treatment, *T-LowThreshold*.

[20]Note that the regressions in Table 6 concern group-level data and not individual data, which is why the number of observations is smaller.

Table 5: Treatment effects on playing Close to Defection

Dependent Variable:	Close to Defection					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
T-OnePlayer	-2.535**	-2.585**	-1.216**	-1.122*	-0.0919***	-0.0834***
	(1.024)	(1.083)	(0.5897)	(0.5889)	(0.0288)	(0.0294)
T-LowThreshold	-0.6224	-0.7173*	-1.027***	-1.069***	-0.0709**	-0.0703**
	(0.4533)	(0.3921)	(0.3612)	(0.3650)	(0.0284)	(0.0270)
Constant	-2.652***	3.297	-1.516***	2.137	0.1196***	0.3959***
	(0.2482)	(2.205)	(0.2322)	(1.382)	(0.0245)	(0.1098)
Controls	No	Yes	No	Yes	No	Yes
Fit statistics						
Observations	1,692	1,692	1,692	1,692	1,692	1,692
R^2					0.03533	0.07386
Pseudo R^2	0.02890	0.08459	0.03888	0.08263		
Wald (joint nullity)	3.6184	3.0670	5.0063	6.0010	5.1168	3.6988

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

five. For all periods in between the coefficients are also significant, as illustrated in the lower panel of Figure 3. The average frequency of reaching the threshold increases by 29-33pp for *T-OnePlayer* and by 25-27pp for *T-LowThreshold* (Columns (5) and (6) of Table 6). Again, transforming the coefficients of the logit regression into marginal effects supports the quantification of the effects. (For *T-OnePlayer* they indicate a slightly larger effect of around 33-39pp.) The effects are large and highly significant and hence fully support Hypotheses H-Reached-One and H-Reached-Low.

5.3 Further treatment effects

The treatment effects that we find above should have implications for further outcome variables of interest, in particular for the evolution of endowments and for maintaining sustainability. Figure 4 illustrates the evolution of endowments and sustainability over time for each treatment. While in the baseline treatment only 10.8% of the groups act sustainably in the first five periods, this fraction increases to 33.9% when there is no strategic interaction and to 21.9% when the threshold is lower. Notice that there is a mechanical effect that *T-LowThreshold* leads to weaker decrease in endowments than the two other treatments for the same contributions (below 50%, as it can be seen in Figure 1). This mechanical effect, which comes on top of the behavioral response to the treatments established above, explains why endowments decrease the least in *T-LowThreshold*. For the binary outcome variable *Sustainability* this mechanical effect is absent.

Table 7 estimates the treatment effects of removing strategic interactions and lowering the threshold on groups' *Sustainability* at the end of period five and on their endowments at period six.[21] Columns (1) and (2) indicate that *T-OnePlayer* and *T-LowThreshold* both have a positive impact on a group's probability of maintaining a sustainable behavior until the end of period five. Columns (3) and (4) reveal an increase in the period 6 endowment by 18-21 points for *T-OnePlayer* and by 27-28 points for *T-LowThreshold*, in comparison to *T-Baseline*. The effects are hence

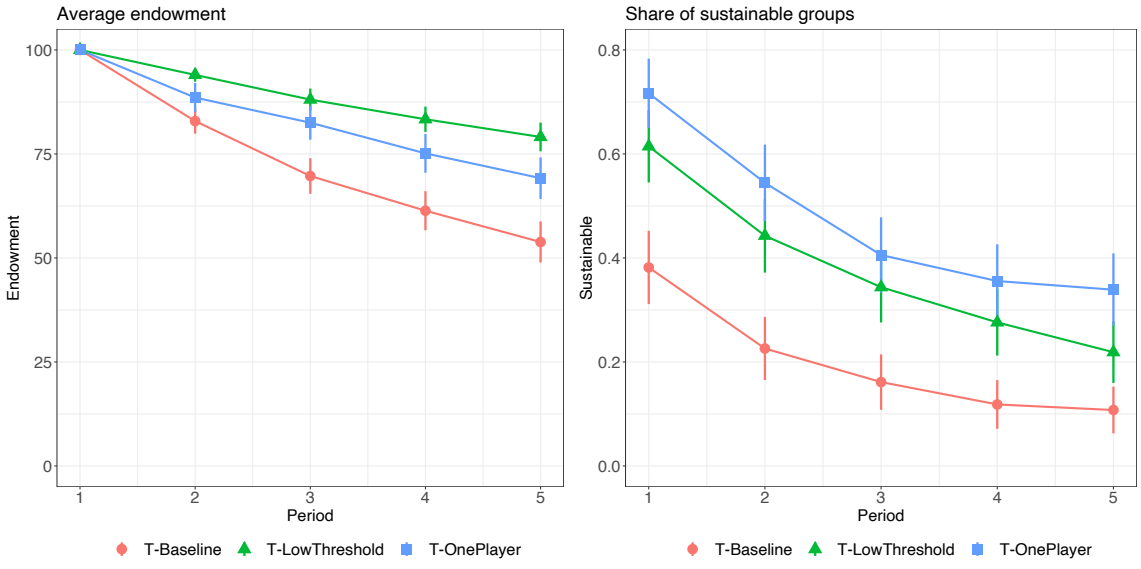
[21] Both variables only depend on decisions in the first five periods.

Table 6: Treatment effects on Threshold Met

Dependent Variable:	Threshold met					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
T-OnePlayer	1.410***	1.686***	1.488***	1.768***	0.2946***	0.3325***
	(0.3595)	(0.3665)	(0.2993)	(0.2995)	(0.0647)	(0.0597)
T-LowThreshold	0.9489***	1.124***	0.9280***	1.088***	0.2478***	0.2711***
	(0.2951)	(0.3217)	(0.2546)	(0.2796)	(0.0521)	(0.0536)
Constant	-0.4823**	2.958*	-0.5051***	4.477**	0.3710***	1.118***
	(0.2133)	(1.707)	(0.1630)	(1.740)	(0.0408)	(0.2725)
Controls	No	Yes	No	Yes	No	Yes
Fit statistics						
Observations	558	558	558	558	558	558
R^2					0.14110	0.19742
Pseudo R^2	0.05860	0.09264	0.06326	0.09412		
Wald (joint nullity)	9.1344	4.0404	14.443	5.5233	14.535	7.5775

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Figure 4: Evolution of Endowments and Sustainability over time, by treatment



Notes: Mean and standard 95% confidence intervals, pooling groups and repetitions of the game. *Sustainability* is a binary variable that equals 1 if a participant's group has reached the sustainability threshold in all periods so far.

substantial. Finally, we estimate treatment effects on the time trend of *Endowment* (Appendix Table B.4). In *T-Baseline*, *Endowment* falls by 11.4 points on average in every period. The decline in *Endowment* is significantly smaller in *T-OnePlayer* (7.5 points) and *T-LowThreshold* (5.3 points). To summarize, we find that that without strategic interaction and with a lower threshold, indeed, endowments remain higher and sustainability can be more often maintained.

Table 7: Treatment effects on Sustainability and Endowment

Dependent Variables:	Sustainability period 5		Endowment period 6	
Model:	(1) Logit	(2) Logit	(3) OLS	(4) OLS
Variables				
T-OnePlayer	1.448***	1.593***	17.66**	21.11***
	(0.5007)	(0.5117)	(6.880)	(6.792)
T-LowThreshold	0.8433*	1.004**	26.71***	28.14***
	(0.4597)	(0.4731)	(5.406)	(5.653)
Constant	-2.116***	1.353	47.34***	91.80***
	(0.4112)	(2.048)	(4.504)	(28.99)
Controls	No	Yes	No	Yes
Fit statistics				
Observations	558	558	558	558
R^2			0.10391	0.15144
Pseudo R^2	0.05001	0.07411		
Wald (joint nullity)	4.2717	1.8266	12.214	4.3837

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

5.4 Test of secondary hypotheses: agreeableness and cognitive ability

Hypothesis H-Agree stipulates a positive relation between *Agreeableness* and *Contribution Share*.[22] In Table 8, we test this hypothesis by regressing *Contribution Share* on *Agreeableness*, treatment dummies and the set of control variables (*Age*, *Female*, *Raven*, *Risk Tolerance*, *Conscientiousness*). Columns (2), (4) and (6) additionally include the interaction of *Agreeableness* with treatment dummies to allow for heterogeneous effects of *Agreeableness* across treatments.

Table 8 indicates that *Agreeableness* does not significantly affect *Contribution Share* in the first period of the game, while there is a significant effect in period five and on the average. A one-point increment in *Agreeableness* raises contribution shares by 0.36pp in period five (Column (3)) and by 0.26pp on average (Column (5)), a small effect.[23] The result suggests that participants who score high on *Agreeableness* maintain slightly higher contribution levels throughout the game.[24]

Columns (4) and (6) indicate that *Agreeableness* has heterogeneous effects across treatments.[25] The effect of *Agreeableness* is strongest in the baseline treatment. The effect is weaker in *T-LowThreshold* indicating that *Agreeableness* matters more in situations where *Cooperation* is more

[22]For treatment effects concerning the outcome variable *Contribution Share*, see Appendix B.5.

[23]The *Agreeableness* score ranges from 12 to 60 points with a standard deviation of 6.38. A participant whose *Agreeableness* is one standard deviation higher than another participant's would therefore contribute on average $6.38 \times 0.26 = 1.7pp$ more.

[24]This observation is in line with the finding that agreeableness is associated with cooperative behavior and with positive reciprocity (e.g. Dohmen et al., 2008; Volk et al., 2011).

[25]Studying interaction effects reduces statistical power (e.g., Maxwell et al., 2017). We use the interaction effects to qualitatively show which treatments drive the overall effect.

difficult, i.e., when it is not an equilibrium of the game. We also observe that the effect of *Agreeableness* disappears in *T-OnePlayer*, indicating that *Agreeableness* only matters when multiple players are involved in the game. Intuitively, participants with higher agreeableness score show more agreeable behavior toward others. Overall, Hypothesis H-Agree finds some support, but mainly in the pure social dilemma situation.

Table 8: Effect of Agreeableness on Contribution Shares (OLS)

Dependent Variable:	Contribution Share					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Agreeableness	0.0003	0.0003	0.0036***	0.0060***	0.0026***	0.0046***
	(0.0013)	(0.0025)	(0.0013)	(0.0019)	(0.0010)	(0.0015)
T-OnePlayer × Agree.		-0.0033		-0.0076**		-0.0069**
		(0.0039)		(0.0035)		(0.0030)
T-LowThreshold × Agree.		0.0009		-0.0040*		-0.0029
		(0.0027)		(0.0023)		(0.0019)
Constant	0.4500***	0.4469***	0.2891***	0.1566	0.3464***	0.2400***
	(0.0793)	(0.1334)	(0.0703)	(0.1029)	(0.0600)	(0.0840)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	1,692	1,692	1,692	1,692	1,692	1,692
R^2	0.19261	0.19396	0.16136	0.16592	0.30834	0.31503
Wald (joint nullity)	28.954	25.023	24.533	26.462	38.345	38.567

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*
Notes: The *Agreeableness* score ranges from 12 to 60 points with a standard deviation of 6.38.

The final hypothesis, Hypothesis H-Raven, predicts that participants with high cognitive abilities, as measured by high *Raven* scores, more often play equilibrium strategies in the Sustainability Game. Tables 9 and 10 test the relation between *Raven* score and playing *Close to Cooperation* and *Close to Defection*, respectively.

Columns (1) and (3) of Table 9 report the effect of *Raven* on the probability of playing *Close to Cooperation* in period one and five, respectively. We control for treatments and the set of other control variables which is *Age*, *Female*, *Risk Tolerance*, *Agreeableness* and *Conscientiousness*. The estimate is positive and highly significant in both periods, thus supporting that more intelligent players more often play the cooperative equilibrium. Column (5) reports the effect of *Raven* on the average frequency of playing *Close to Cooperation* in the first five periods of the game. It indicates that a one-point increase in *Raven* is associated with a 3pp increase in the frequency of playing *Close to Cooperation*. It follows that a participant whose *Raven* score is one standard deviation higher than another participant's would cooperate on average $3.11 \times 0.03 = 9pp$ more often.

However, it is important to remember that *Close to Cooperation* is not an equilibrium strategy in *T-Baseline*, but only in *T-OnePlayer* and *T-LowThreshold*. According to Hypothesis H-Raven, the effect of cognitive ability should therefore be stronger in *T-OnePlayer* and *T-LowThreshold* relative to *T-Baseline*. Columns (2), (4), and (6) examine the heterogeneity of the effect of *Raven* across treatments. In the baseline treatment *T-Baseline* the effect is weaker than overall. On the other hand, we observe a weakly significant increase in the effect of *Raven* on playing *Close to Cooperation* in *T-LowThreshold* relative to *T-Baseline* in period one and on average. Therefore,

the effect of *Raven* on playing *Close to Cooperation* is mainly driven by *T-LowThreshold* where *Cooperation* is indeed an equilibrium. This observation is in line with the causal evidence provided by Proto et al. (2019) who show that participants with high *Raven* score are better able to select the Pareto-dominant equilibrium.

Table 9: Effect of Raven on playing Close to Cooperation

Dependent Variable:	Close to Cooperation					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
Raven	0.2058***	0.1165*	0.1317***	0.0717	0.0302***	0.0150
	(0.0434)	(0.0639)	(0.0344)	(0.0592)	(0.0063)	(0.0098)
T-OnePlayer × Raven		0.1420		0.0108		0.0241
		(0.1158)		(0.0993)		(0.0178)
T-LowThreshold × Raven		0.1587*		0.1126		0.0256*
		(0.0899)		(0.0795)		(0.0135)
Constant	-1.676	-0.8863	-1.432*	-0.8301	0.1412	0.2609*
	(1.045)	(1.044)	(0.8541)	(0.8730)	(0.1545)	(0.1471)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	1,692	1,692	1,692	1,692	1,692	1,692
R^2					0.13610	0.14817
Pseudo R^2	0.09267	0.10015	0.07353	0.07753		
Wald (joint nullity)	5.3471	6.0530	4.8813	4.3861	7.1612	7.4095

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: The *Raven* score ranges from 0 to 12 with a standard deviation of 3.11.

Concerning *Defection*, Columns (1), (3) and (5) of Table 10 indicate no significant influence of *Raven* on the probability of playing *Close to Defection*. According to Hypothesis H-Raven though, we only expect *Raven* to have an impact on playing *Close to Defection* when *Defection* is an equilibrium strategy, i.e., in treatments *T-Baseline* and *T-LowThreshold*. For *T-Baseline* Columns (2), (4) and (6) show a significant positive effect in period five and on average, but not in period one. This suggests that high cognitive ability players do not necessarily play the unique equilibrium strategy more often from the beginning of the game, but settle for it more frequently later. Hence, if *Raven* positively affects the probability of *Close to Defection*, this is mainly driven by the treatment where *Defection* is the unique equilibrium strategy. Interestingly, in the treatment *T-LowThreshold*, in which both *Defection* and *Cooperation* are equilibria, this effect is nullified, again consistent with the idea that in a situation of equilibrium selection, high *Raven* participants might more often select the Pareto-efficient equilibrium.

5.5 Potential learning effects

Our analysis is based on pooling six repetitions of the game and shown to be robust when focusing on the first game. To explore potential learning effects, we have re-run our main regressions and graphs separated for each repetition of the game, which range from 1 (first game) to 6 (last game). Several insights emerge. Most of them can be seen in Figure 5, which shows the main outcome variables by repetition of the game, plus the average contribution shares.

Table 10: Effect of Raven on playing Close to Defection

Dependent Variable:	Close to Defection					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
Raven	-0.0486	-0.0554	0.0727	0.1171**	0.0029	0.0101*
	(0.0615)	(0.0671)	(0.0488)	(0.0529)	(0.0035)	(0.0059)
T-OnePlayer × Raven		0.1547		0.0096		-0.0093
		(0.1007)		(0.1668)		(0.0074)
T-LowThreshold × Raven		0.0100		-0.1340		-0.0127*
		(0.1346)		(0.1051)		(0.0072)
Constant	3.297	3.368	2.137	1.644	0.3959***	0.3365***
	(2.205)	(2.071)	(1.382)	(1.359)	(0.1098)	(0.1046)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	1,692	1,692	1,692	1,692	1,692	1,692
R^2					0.07386	0.08242
Pseudo R^2	0.08459	0.08511	0.08263	0.08750		
Wald (joint nullity)	3.0670	2.9632	6.0010	5.5413	3.6988	3.2550

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

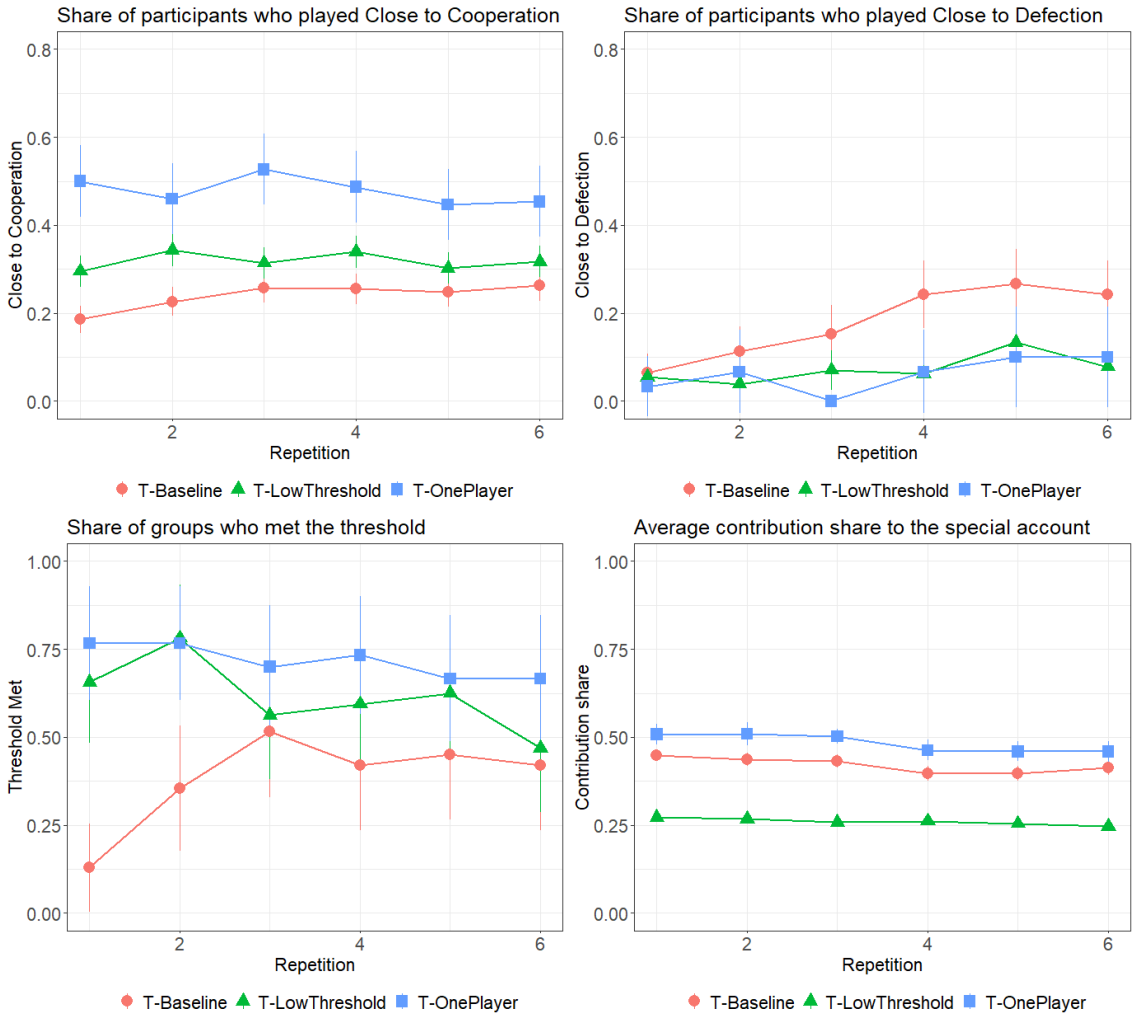
First, learning effects are limited. There are no drastic changes over time and the treatment effects (of *T-OnePlayer* or *T-LowThreshold* versus *T-Baseline*) have the same sign in any repetition for any outcome variable. Second, in *T-Baseline* participants increasingly often play not only close to the equilibrium strategy *Defection*, but also to the efficient strategy *Cooperation*. As a consequence, the threshold is reached more often in later repetitions of the game. Third, treatment effects are strong and significant in the first repetition of the game (as we already partially knew from the robustness tests conducted in Appendix Tables B.1-B.3). Treatment effects tend to get weaker with repetitions. For instance, the treatment effect of *T-OnePlayer* on *Close to Cooperation* shrinks from $1.480^{\star\star\star}$ in the first game to $0.8437^{\star\star}$ in the last repetition of the game. The reason is not necessarily a change over time in *T-OnePlayer*, but rather the change in *T-Baseline*, which narrows the gap between treatments.

Fourth, subjects do not increase their payoffs over time (not illustrated) and some noisy behavior persists. In particular, in *T-OnePlayer* about half of the times subjects do not contribute close to the threshold ("*Cooperation*"), the unique and efficient equilibrium. This share does not significantly change with repetitions of the game. Moreover, as the histogram of contributions (Appendix Figure B.1) reveals, there are also a few subjects contributing above 50%, a dominated strategy in *T-OnePlayer*. In *T-Baseline* the effect of playing *Cooperation* and reaching the threshold more often might be offset by the higher frequency of *Defection*, as payoffs do not significantly change in any treatment.

5.6 Further results: waste and environmental attitudes

The test of the secondary hypotheses shows that there are associations between the personality traits agreeableness and cognitive ability with behavior in the Sustainability Game. In particular,

Figure 5: Learning effects: main variables evolution over repetitions, by treatment



Notes: Mean and standard 95% confidence intervals, pooling groups and first five rounds of the game. *Close to Cooperation* and *Close to Defection* are binary variables that equal 1 if a participant has submitted a contribution share within a $\pm 2pp$ band around the *Cooperation* and *Defection* strategies, respectively. *Threshold met* is a binary variable that equals 1 if a participant's group has reached the sustainability threshold.

in line with Hypothesis H-Raven, players with high cognitive ability more often play equilibrium strategies. As a corollary, we expect that they are less likely to choose "wasteful" strategies. *Waste* occurs when a group contributes to the special account more than the sustainability threshold. In Appendix B.6 we explore this relation between cognitive ability and *Waste*. Participants with higher Raven score indeed less often contribute above the threshold z (Appendix Table B.6) with the consequence that groups with higher average Raven score less often induce waste (Appendix Table B.7).[26]

Finally, we explore how cooperative behavior in the Sustainability Game correlates with environmental attitudes. We find systematic correlations with the New Ecological Paradigm (NEP) score, but no relation with our small CO2 footprint score. The NEP measures the endorsement of a "pro-ecological" world view (Dunlap et al., 2000).[27] The results in Appendix Table B.7 indicate that participants with a stronger ecological concern are more likely to play *Close to Cooperation* beyond the effects of other control variables (Table B.8). The effect is significant overall, but seems to be mainly driven by *T-OnePlayer*. Appendix Table B.9 further shows that a higher NEP score is associated with a lower probability of playing *Close to Defection*, without any clear differences among the treatments. Therefore, sustainable behavior in our Sustainability Game is positively associated with pro-environmental orientation.

6 Discussion

Sustainability is defined as using resources today in a way that does not compromise the availability of resources tomorrow. This definition entails that there is a threshold that distinguishes sustainable behavior, which can be repeated infinitely, from unsustainable behavior, which cannot. In this paper we propose a model of sustainability. It differs from the standard public goods (PG) and common pool resource (CPR) games by being dynamic and having a threshold. Both these aspects are necessary to capture the meaning of sustainability, as defined above. In particular, our model incorporates a resource's capacity to absorb a certain level of consumption, while over-consumption (irreversibly) diminishes the resource. This feature further differentiates our model from most of the pre-existing literature and makes conservation of personal health or of the environment ideal applications.

Theoretically and experimentally, we find that cooperation is fostered by excluding interaction and by lowering the sustainability threshold. The former result reflects analogous fundamental findings for PG and CPR games, where making one player solely involved eliminates the free-rider problem. The latter result goes beyond a mechanical effect of reaching a threshold more often just because it is set lower. We find in particular that lowering the threshold makes people choose zero contributions less frequently. Drawing conclusions for the goal of sustainability (be it in the environmental domain or concerning personal health), our results suggests that technological progress need not be seen as a substitute to behavior change: If it lowers the threshold, it can also work as a complement that fosters behavior change toward sustainability.[28]

[26] Moreover, there are some learning effects as subjects over-contribute less often in later repetitions of the game. This trend is strongest in *T-LowThreshold*, where the share of groups who contribute strictly above the threshold (*Waste*) even drops by more than 20pp. Still, in the first and in the last games it is the case that higher Raven scores significantly correlate with less over-contribution and wasteful behavior.

[27] The NEP score ranges from 26 to 71 with mean 55.5 and standard deviation 7.7.

[28] This is even more the case for other types of technology, which improve the infrastructure for sustainable behavior (e.g., bike lanes), or which support sustainable behavior directly (e.g., apps that meter health related activities).

A Appendix: Proofs

A.1 Proof of Proposition 1: social planner's solution

Proposition A.1 below nests Proposition 1 of the main text as a special case.

Proposition A.1 (Social Planner's Solution, General)**.** *Consider a social planner that maximizes the aggregate expected payoff* $EP_0 := \sum_{t=0}^{\infty} \delta^t \left(ne_t - c_t ne_t\right)$ *over the fraction* c_t *of the total endowment that the group contributes.*

1. *If* $\delta > \underline{\delta} := \frac{1}{1+ng(1-z)}$*, the social planner chooses* $c_t^* = z$ *at each period* t*.*
2. *If* $\delta = \underline{\delta}$*, the social planner chooses* $c_t^* \in [0, z]$*.*
3. *If* $\delta < \underline{\delta}$*, the social planner chooses* $c_t^* = 0$*.*

An aggregate contribution above the threshold is wasteful, which means that the social planner chooses the aggregate contribution share c_t such that $c_t \leq z \ \forall t$. It follows from Equation 1 that $e_{t+1} = e_t - g\left(zne_t - c_t e_t\right) \ \forall t$. To find the social planner's solution, we can therefore maximize the value $V(e_t)$ depending on consumption in t and the discounted future value $V(e_{t+1})$ in the following way:

$$\begin{aligned} \max_{c_t} V(e_t) = ne_t - c_t ne_t + \delta V(e_{t+1}) & \\ \text{s.t.} & \\ c_t \leq z & \\ c_t \geq 0 & \\ e_{t+1} = e_t - g(zne_t - c_t ne_t)\,. & \end{aligned} \tag{A.1}$$

Maximizing the corresponding Lagrangian

$$\max_{c_t} V(e_t) = ne_t - c_t ne_t + \delta V(e_t - gzne_t + gc_t ne_t) - \lambda_{1,t}(c_t - z) + \lambda_{2,t} c_t,$$

we get a first order condition w.r.t. c_t (the FOC), an envelope condition capturing the derivative of the value function with respect to the endowment (the EC) and two complementary slackness conditions:

$$\begin{aligned} \text{(FOC)} \qquad -ne_t + \delta V'(e_{t+1}) gne_t - \lambda_{1,t} + \lambda_{2,t} &= 0 \\ \text{(EC)} \qquad V'(e_t) &= n - nc_t + \delta V'(e_{t+1})(1 - gnz + gnc_t) \\ \lambda_{1,t}(c_t - z) &= 0 \\ \lambda_{2,t} c_t &= 0 \\ c_t - z &\leq 0 \\ c_t &\geq 0 \\ \lambda_{1,t}, \lambda_{2,t} &\geq 0\,. \end{aligned} \tag{A.2}$$

We distinguish three cases. First, suppose that the condition $c_t \leq z$ is binding. In this case, $c_t = z$, and $\lambda_{2,t} = 0$. Moreover, for $c_t = z$ it holds that $e_{t+1} = e_t$. The FOC and EC in System of Equations A.2 become $\delta V'(e_t) gne_t = ne_t + \lambda_{1,t}$ and $V'(e_t) = n - nz + \delta V'(e_t)$. Solving for $\lambda_{1,t}$ and $V'(e_t)$ yields $V'(e_t) = \frac{n(1-z)}{1-\delta}$ and $\lambda_{1,t} = \frac{\delta g n^2 e_t (1-z)}{1-\delta} - ne_t$. The condition $\lambda_{1,t} \geq 0$ is then equivalent to $\delta \geq \frac{1}{1+ng(1-z)}$. Hence, $c_t = z$ is a solution iff $\delta \geq \frac{1}{1+ng(1-z)} (= \underline{\delta})$.

Second, suppose that the condition $c_t \geq 0$ is binding, and in consequence that $c_t = 0$, $\lambda_{1,t} = 0$. It follows for the FOC and the EC in System of Equations A.2 that $-ne_t + \delta V'(e_{t+1}) gne_t + \lambda_{2,t} = 0$ and $V'(e_t) = n + \delta V'(e_{t+1})(1 - gnz)$. We guess and verify that the value function $V(e_t)$ is linear in e_t: $V(e_t) = ke_t$, with k being a positive constant. [29] This implies $V'(e_t) = V'(e_{t+1}) = k$. We can then use the EC to determine that $k = V'(e_t) = \frac{n}{1-\delta+\delta gnz}$. Using the FOC we determine that $\lambda_{2,t} = ne_t - \frac{\delta g n^2 e_t}{1-\delta+\delta gnz}$. The condition $\lambda_{2,t} \geq 0$ is then equivalent to $\delta \leq \frac{1}{1+ng(1-z)} (= \underline{\delta})$. Thus, $c_t = 0$ is a viable solution to our problem iff $\delta \leq \underline{\delta}$.

[29] For the 'guess and verify' technique see, e.g., Chapter 3 of Ljungqvist and Sargent (2018): If we find an expression for k which solves all the maximizing conditions, we find the value function.

Third, consider an interior solution $c_t \in [0, z]$. This is the case if $\lambda_{1,t} = 0$ and $\lambda_{2,t} = 0$, which leads to the following FOC and EC: $\delta V'(e_{t+1})ge_t n = e_t n$ and $V'(e_t) = n(1 - c_t) + \delta V'(e_{t+1})(1 - gzn + gc_t n)$. The FOC implies that $V'(e_t) = \frac{1}{\delta g}$, $\forall t$, such that the EC holds independently of c_t. It follows for $\delta ng(1 - z) + \delta = 1$, or equivalently for $\delta = \underline{\delta}$, that any contribution level above or equal to zero and below or equal to the sustainability threshold z is optimal. These three results together establish Proposition A.1.

A.2 Proof of Proposition 2: Markov Perfect Equilibria

We prove Proposition 2 here. For $n = 1$ the solution of the social planner's problem, Proposition A.1, applies. Observing that for $n = 1$, we have $\underline{\delta} = \bar{\delta} = \bar{\bar{\delta}}$, it follows that a single player chooses *Defection* for $\delta < \bar{\delta}$; and *Cooperation* for $\delta > \bar{\delta}$, as claimed by Proposition 2. Suppose from now on that $n > 1$.

We first show that there is no MPE with waste. Assume the contrary. Then there is a time t at which the sum of contributions is above the threshold, i.e., $\sum_{i=1}^{n} c_{i,t}e_t > Z_t$. A single player l could deviate by slightly reducing her contribution to $c'_{l,t} = c_{l,t} - \varepsilon$, with $\varepsilon > 0$ and $\varepsilon e_t \leq \sum_{i=1}^{n} c_{i,t}e_t - Z_t$. Her benefits in time t increase, while the state variable e_{t+1} stays constant. As Markov strategies only depend on the state variable, the continuation of the game is unchanged. Hence, reducing her contribution is a strict improvement for player l, in contradiction to the assumption of the original situation being a MPE. Moreover, using the same argument, if there is a beneficial deviation that yields waste, there is a more attractive deviation that does not exceed the aggregate threshold.

Player i maximizes (2), taking the contributions of the others as given. Let $\bar{c}_t$ denote the average contribution share of the other players. Using that the overall contributions never exceed the threshold, player i maximizes

$$\begin{aligned} \max_{c_{i,t}} V(e_t) &= e_t - c_{i,t}e_t + \delta V(e_{t+1}) \\ &\text{s.t.} \\ c_{i,t}e_t + (n-1)\bar{c}_t e_t &\leq zne_t \\ c_{i,t} &\geq 0 \\ e_{t+1} &= e_t - g(zne_t - c_{i,t}e_t - (n-1)\bar{c}_t e_t)\,. \end{aligned} \tag{A.3}$$

Taking derivatives of the corresponding Lagrangian

$$\begin{aligned} \max_{c_{i,t}} V(e_t) = e_t - c_{i,t}e_t + \delta V(e_t - gzne_t + gc_{i,t}e_t + g(n-1)\bar{c}_t e_t) \\ -\lambda_{1,t}e_t(c_{i,t} + (n-1)\bar{c}_t - zn) + \lambda_{2,t}c_{i,t} \end{aligned}$$

leads to the following first order condition (FOC), envelope condition (EC) and two complementary slackness conditions:

$$\begin{aligned} \text{(FOC)} \qquad -e_t + \delta V'(e_{t+1})ge_t - \lambda_{1,t} + \lambda_{2,t} &= 0 \\ \text{(EC)} \qquad V'(e_t) &= 1 - c_{i,t} + \delta V'(e_{t+1})(1 - gzn + gc_{i,t} + g\bar{c}_t(n-1)) \\ \lambda_{1,t}e_t(c_{i,t} + (n-1)\bar{c}_t - zn) &= 0 \\ \lambda_{2,t}c_{i,t} &= 0 \\ c_{i,t} + (n-1)\bar{c}_t - zn &\leq 0 \\ c_{i,t} &\geq 0 \\ \lambda_{1,t}, \lambda_{2,t} &\geq 0. \end{aligned} \tag{A.4}$$

If the endowments are zero, $e_t = 0$, then contributions in all further periods are zero for any strategy profile. Hence, we focus on $e_t > 0$. Then the first complementary slackness condition simplifies to $\lambda_{1,t}(c_{i,t} + (n-1)\bar{c}_t - zn) = 0$. We search for symmetric equilibria, meaning that all players contribute the same fraction of their endowment

First, suppose that $c_{i,t} \geq 0$ is binding. In this case it holds that $c_{i,t} = 0$ and the first complementary slackness condition further simplifies to $\lambda_{1,t}((n-1)\bar{c}_t - zn) = 0$. This is satisfied either if $(n-1)\bar{c}_t = zn$ or $\lambda_{1,t} = 0$. The former subcase cannot constitute a symmetric equilibrium, as $c_{i,t} = 0$ while $\bar{c}_t = \frac{nz}{n-1} > 0$. Consider hence the latter subcase: $\lambda_{1,t} = 0$. It follows for the FOC

and the EC that:

$$\begin{aligned} \text{(FOC)} \qquad & -e_t + \delta V'(e_{t+1}) g e_t + \lambda_{2,t} = 0 \\ \text{(EC)} \qquad & V'(e_t) = 1 + \delta V'(e_{t+1})\big(1 - gzn + g\bar{c}_t(n-1)\big). \end{aligned} \tag{A.5}$$

Again, we guess and verify that the value function $V(e_t)$ is linear in e_t: $V(e_t) = ke_t$, with k being a positive constant. This implies $V'(e_t) = V'(e_{t+1}) = k$. We can then use the EC in System of Equations A.5 to determine k and the FOC to determine $\lambda_{2,t}$:

$$\begin{aligned} k = V'(e_t) &= \frac{1}{1 - \delta + \delta g z n}, \\ \lambda_{2,t} &= \left(1 - \frac{\delta g}{1 - \delta + \delta g z n}\right) e_t. \end{aligned} \tag{A.6}$$

With $V(e_t) = \frac{1}{1-\delta+\delta g z n} e_t$ we found a consistent solution for the value function that satisfies the FOC and the EC. The condition $\lambda_{2,t} \geq 0$ is then equivalent to $\delta g(1 - zn) \leq 1 - \delta$. Thus, $c_t = 0$ is a symmetric MPE iff $\delta \leq \frac{1}{1+g(1-zn)} (= \bar{\bar{\delta}})$.

Second, suppose that $c_{i,t} e_t \leq nz e_t - \bar{c}_t(n-1)e_t$ is binding. In this case it holds that $c_{i,t} = nz - \bar{c}_t(n-1)$. The second complementary slackness condition, $\lambda_{2,t} c_{i,t} = 0$, is satisfied either if $c_{i,t} = 0$ or if $\lambda_{2,t} = 0$. The former subcase cannot constitute a symmetric equilibrium, as $c_{i,t} = 0$ while $\bar{c}_t(n-1) + c_{i,t} = nz$ implies $\bar{c}_t = \frac{nz}{n-1} > 0$. Consider hence the latter subcase: $\lambda_{2,t} = 0$. It follows for the FOC and the EC that

$$\begin{aligned} \text{(FOC)} \qquad & -e_t + \delta V'(e_{t+1}) g e_t - \lambda_{1,t} = 0 \\ \text{(EC)} \qquad & V'(e_t) = 1 - zn + \bar{c}_t(n-1) + \delta V'(e_{t+1}). \end{aligned} \tag{A.7}$$

For $c_{i,t} = nz - \bar{c}_t(n-1)$, endowments are constant, such that $e_{t+1} = e_t$. We can solve the System of Equations A.7 for $\lambda_{1,t}$ and $V'(e_t)$ using that it follows from symmetry and $c_{i,t} = nz - \bar{c}_t(n-1)$ that $c_{i,t} = \bar{c}_t = z$:

$$\begin{aligned} V'(e_t) &= \frac{1 - z}{1 - \delta} \\ \lambda_{1,t} &= \left(\frac{\delta g(1-z)}{1-\delta} - 1\right) e_t \,. \end{aligned} \tag{A.8}$$

The condition $\lambda_{1,t} \geq 0$ is then equivalent to $\delta \geq \frac{1}{1+g(1-z)} (= \bar{\delta})$. Thus, $c_t = z$ is a symmetric MPE iff $\delta \geq \bar{\delta}$.

Third, suppose that we have an interior solution, in which case it holds that $\lambda_{1,t} = 0$ and $\lambda_{2,t} = 0$. The FOC and EC from System of Equations A.5 become

$$\begin{aligned} \text{(FOC)} \qquad & \delta V'(e_{t+1}) g e_t = e_t, \\ \text{(EC)} \qquad & V'(e_t) = 1 - c_{i,t} + \delta V'(e_{t+1})\big(1 - gzn + gc_{i,t} + g\bar{c}_t(n-1)\big). \end{aligned} \tag{A.9}$$

It follows from the FOC that $V'(e_{t+1}) = \frac{1}{\delta g} \, \forall t$. Plugging this result into the EC, while using symmetry in the sense that $c_{i,t} = \bar{c}_t = \tilde{c}$, we find:

$$\tilde{c} = \frac{1 - \delta + \delta g(zn - 1)}{\delta g(n-1)}. \tag{A.10}$$

These results together establish Proposition 2. Notice, however, that the *Inbetween* strategy profile derived last is only a weak equilibrium: If the other players all contribute $\tilde{c}$, player i is indifferent with respect to her own contribution (up to a level where she starts to induce waste). We can see this by replacing $\bar{c}_t$ by $\tilde{c}$ in EC A.9. After some simplifications, we receive $V'(e_t) = 1 - c_{i,t} + V'(e_{t+1})\big(\delta g c_{i,t} + 1 - \delta g\big)$. Now, using $V'(e_t) = V'(e_{t+1}) = \frac{1}{\delta g}$, we observe that the equation holds independently of the level of $c_{i,t}$.

B Appendix: Additional Figures and Tables

B.1 Attrition

Some potential participants did not complete the study. Most of them dropped out at the very beginning, before the game started. In fact, 98 subjects, i.e., 21.1% of all 464 potential participants, dropped out early on. Further 39 subjects, i.e., 8.4% of all, could not enter the game because they could not be matched in a group of four. Only twelve subjects dropped out after the start of the game. This, however, affected 33 further subjects who were their group members. Together, this yields 45 subjects, i.e., 9.7% of all potential participants, who were lost during the game.

In terms of groups, 105 groups started the game and 93, i.e., 88.6%, completed it. The twelve groups that dropped out are distributed as follows: seven groups in *T-Baseline*, one in *T-OnePlayer* and four in *T-LowThreshold*. We observe most dropouts in *T-Baseline*, less in *T-LowThreshold*, and the least in *T-OnePlayer*. Clearly, the likelihood that a group drops out is higher in the two treatments with $n = 4$ than in the treatment with $n = 1$, but we also observe that more groups dropped out in *T-Baseline* than in *T-LowThreshold*.

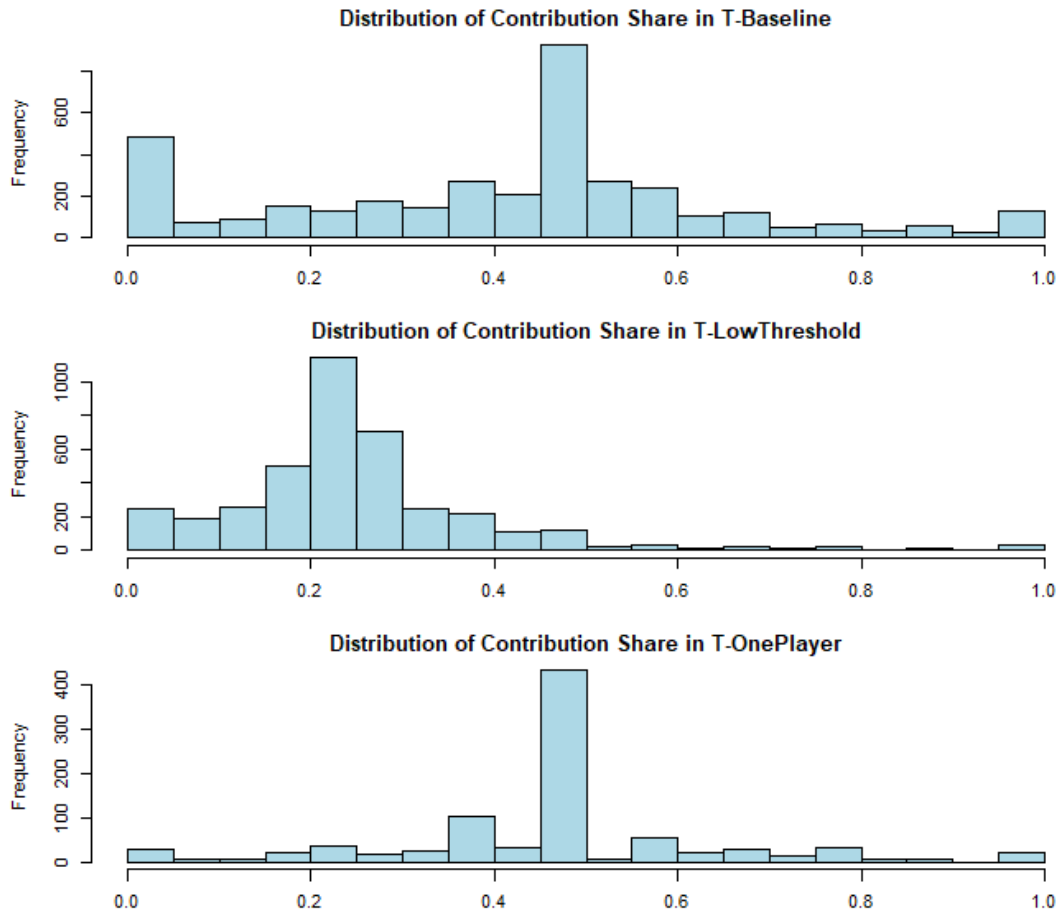
We control for potential attrition effects by using control variables in our regressions. Moreover, comparing the descriptive statistics across treatments, we found no systematic differences. Despite attrition, the characteristics across treatments are balanced.

B.2 Distribution of contribution shares

Figure B.1 shows the distribution of contribution shares for each treatment, pooled over the first five periods. The graphs reveal that the theoretically derived focal strategies are indeed pre-dominant.

In *T-Baseline* the focal strategies are *Cooperation*, which is efficient and means contributing 50%; and *Defection*, which is the unique equilibrium and means contributing 0%. In *T-LowThreshold* the focal strategies are *Cooperation*, which is efficient, an equilibrium, and means contributing 25%; and *Defection*, which is an equilibrium and means contributing 0%. Besides, there is the equilibrium *Inbetween* which means contributing approximately 17.95%. In *T-OnePlayer* there is only one focal strategy: "*Cooperation*," which is efficient and means contributing 50%.

Figure B.1: Histograms of contribution shares pooled over first five periods, by treatment



B.3 Robustness test: first game only

Table B.1: Treatment effects on playing Close to Cooperation (first game only)

Dependent Variable:	Close to Cooperation					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
T-OnePlayer	1.480***	1.792***	2.178***	2.409***	0.3145***	0.3284***
	(0.4176)	(0.4739)	(0.4809)	(0.5118)	(0.0716)	(0.0643)
T-LowThreshold	-0.1241	-0.1664	0.9323**	0.9915**	0.1098**	0.1077**
	(0.2782)	(0.2899)	(0.3802)	(0.3993)	(0.0475)	(0.0458)
Constant	-0.9333***	-4.118**	-1.910***	-2.610	0.1855***	-0.0513
	(0.1680)	(1.704)	(0.3050)	(1.634)	(0.0285)	(0.1737)
Controls	No	Yes	No	Yes	No	Yes
Fit statistics						
Observations	282	282	282	282	282	282
R^2					0.08825	0.19977
Pseudo R^2	0.04412	0.14571	0.08006	0.13015	0.16457	0.39698
Wald (joint nullity)	7.1876	6.1437	10.271	3.6140	10.485	7.9443

Clustered (Group) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table B.2: Treatment effects on playing Close to Defection (first game only)

Dependent Variable:	Close to Defection					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
T-OnePlayer	-15.75***	-16.51***	-0.6931	-0.4356	-0.0482***	-0.0414***
	(1.001)	(1.065)	(1.094)	(1.068)	(0.0133)	(0.0137)
T-LowThreshold	0.6690	0.4704	-0.1757	-0.1343	-0.0251	-0.0225
	(1.222)	(1.549)	(0.5638)	(0.5232)	(0.0178)	(0.0164)
Constant	-4.812***	6.492	-2.674***	4.062	0.0548***	0.2911***
	(1.001)	(5.192)	(0.3798)	(3.289)	(0.0116)	(0.1043)
Controls	No	Yes	No	Yes	No	Yes
Fit statistics						
Observations	282	282	282	282	282	282
R^2					0.01806	0.10140
Pseudo R^2	0.02983	0.39067	0.00409	0.17393	-0.01300	-0.07630
Wald (joint nullity)	397.74	656.95	0.21476	2.4504	6.8059	3.4844

Clustered (Group) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table B.3: Treatment effects on Threshold Met (first game only)

Dependent Variable:	Threshold met					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
T-OnePlayer	3.099***	4.036***	2.442***	2.893***	0.3778***	0.4137***
	(0.6994)	(0.8380)	(0.6240)	(0.7950)	(0.0720)	(0.0655)
T-LowThreshold	2.556***	2.876***	1.994***	2.011***	0.3708***	0.3621***
	(0.6631)	(0.7451)	(0.5778)	(0.6392)	(0.0597)	(0.0683)
Constant	-1.910***	-1.215	-1.056**	-0.1138	0.3355***	0.8449**
	(0.5446)	(3.042)	(0.4172)	(3.257)	(0.0476)	(0.3533)
Controls	No	Yes	No	Yes	No	Yes
Fit statistics						
Observations	93	93	93	93	93	93
R^2					0.32822	0.39260
Pseudo R^2	0.24224	0.32997	0.17765	0.22325	0.82506	1.0340
Wald (joint nullity)	10.699	3.4754	9.2501	2.2215	21.986	9.1482

Clustered (Group) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

B.4 Evolution of endowments

Table B.4: Treatment effects and endowment time trend

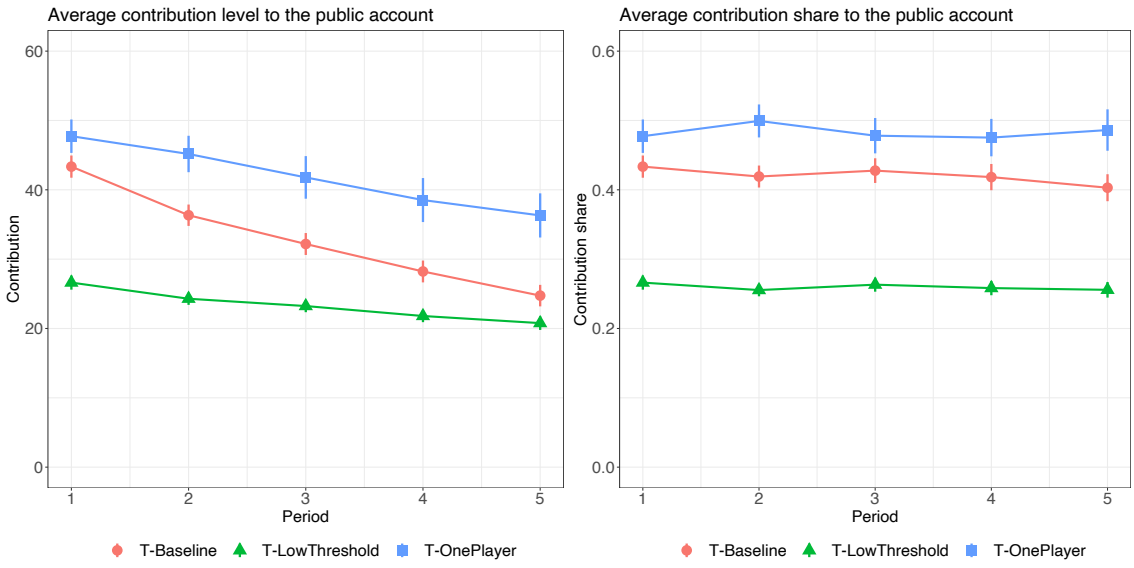
Dependent Variable:	Endowment	
Model:	(1)	(2)
Variables		
Period	-11.39***	-11.39***
	(1.144)	(1.146)
T-OnePlayer × Period	3.883**	3.883**
	(1.616)	(1.618)
T-LowThreshold × Period	6.139***	6.139***
	(1.311)	(1.313)
T-OnePlayer	-2.129	0.2379
	(1.443)	(2.505)
T-LowThreshold	-3.076***	-2.020
	(1.117)	(1.808)
Constant	107.7***	135.3***
	(1.007)	(17.71)
Controls	No	Yes
Fit statistics		
Observations	2,790	2,790
R^2	0.24353	0.28172
Wald (joint nullity)	42.612	21.592

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

B.5 Evolution of contributions

Figure B.2 shows the levels and shares of contributions over time, separated by treatment.

Figure B.2: Contribution to the special account (levels and shares)



Notes: Mean and standard 95% confidence intervals, pooling groups and repetitions of the game

Table B.5 estimates the effect of *T-OnePlayer* and *T-LowThreshold* on the contribution shares to the special account. The dependent variable is considered as continuous, which allows us to run OLS regressions in all columns. Removing strategic interactions increases contribution shares by about 4pp in period one and 8pp in period five. The effect is significant and amounts to about 6pp on average for all periods. The result is consistent with Hypothesis H-CoopDef-One: if more players choose *Cooperation* and fewer choose *Defection*, we can expect an increase in average contribution shares. The results also indicate that lowering the threshold reduces contribution shares by about 16pp. When comparing contribution shares in *T-LowThreshold* and *T-Baseline* there are two opposing effects at play. On the one hand, *T-LowThreshold* should increase average contribution shares, since fewer players choose *Defection*. On the other hand, we can expect a decrease of contribution shares since the socially efficient share is lower. Both effects seem to play a role, as we observe a decline in average contribution shares of 16pp that is smaller than the reduction in the efficient contribution share of 25pp.

Table B.5: Treatment effects on Contribution Shares (OLS)

Dependent Variable:	Contribution Share					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
T-OnePlayer	0.0439*	0.0448*	0.0832***	0.0752***	0.0629***	0.0600***
	(0.0258)	(0.0254)	(0.0278)	(0.0274)	(0.0224)	(0.0223)
T-LowThreshold	-0.1673***	-0.1663***	-0.1473***	-0.1523***	-0.1606***	-0.1631***
	(0.0165)	(0.0165)	(0.0175)	(0.0165)	(0.0146)	(0.0139)
Constant	0.4335***	0.4500***	0.4030***	0.2891***	0.4203***	0.3464***
	(0.0153)	(0.0793)	(0.0156)	(0.0703)	(0.0137)	(0.0600)
Controls	No	Yes	No	Yes	No	Yes
Fit statistics						
Observations	1,692	1,692	1,692	1,692	1,692	1,692
R^2	0.18750	0.19261	0.13113	0.16136	0.28520	0.30834
Wald (joint nullity)	89.560	28.954	70.348	24.533	122.59	38.345

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

B.6 Cognitive ability and wasteful behavior

Table B.6: Effect of Raven on Overcontribution dummy

Dependent Variable:	Overcontribution dummy					
	period 1		period 5		average 1-5	
Model:	(1) Logit	(2) Logit	(3) Logit	(4) Logit	(5) OLS	(6) OLS
Variables						
Raven	-0.1579***	-0.0870	-0.1050***	-0.0761**	-0.0222***	-0.0156**
	(0.0314)	(0.0545)	(0.0240)	(0.0362)	(0.0043)	(0.0063)
T-OnePlayer × Raven		-0.0904		0.0095		-0.0031
		(0.1147)		(0.0978)		(0.0126)
T-LowThreshold × Raven		-0.1194*		-0.0560		-0.0127
		(0.0687)		(0.0505)		(0.0093)
Constant	-1.270	-1.779**	-2.354***	-2.623***	0.1243	0.0654
	(0.8301)	(0.8641)	(0.6772)	(0.7219)	(0.1300)	(0.1373)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	1,692	1,692	1,692	1,692	1,692	1,692
R^2					0.10152	0.10464
Pseudo R^2	0.06042	0.06506	0.05482	0.05608		
Wald (joint nullity)	4.5983	5.0249	7.5497	6.1891	8.6095	7.4970

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Overcontribution is defined as 1 if the participant's contribution share $c_{i,t}$ is strictly larger than the sustainability threshold parameter z and 0 otherwise.

Table B.7: Effect of Raven on Waste dummy

Dependent Variable:	Waste dummy					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
Raven	-0.1308*	-0.1656	-0.1286*	-0.2310**	-0.0251***	-0.0363*
	(0.0737)	(0.1224)	(0.0715)	(0.1083)	(0.0094)	(0.0185)
T-OnePlayer × Raven		-0.0669		0.1338		0.0131
		(0.1662)		(0.1441)		(0.0205)
T-LowThreshold × Raven		0.2186		0.1293		0.0186
		(0.1825)		(0.1831)		(0.0331)
Constant	-1.633	-1.385	-1.057	-0.3703	0.3355	0.4138
	(1.784)	(1.930)	(1.829)	(1.884)	(0.2736)	(0.2962)
Treatmentdummies	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	558	558	558	558	558	558
R^2					0.18387	0.18565
Pseudo R^2	0.10916	0.11786	0.06902	0.07094		
Wald (joint nullity)	3.7497	3.0584	3.2101	2.9939	6.7304	6.1179

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Waste is defined as 1 if the group's aggregate contribution $\sum_{i=1}^{n} c_{i,t} e_t$ is strictly larger than the sustainability threshold Z_t and 0 otherwise.

B.7 Association with New Ecological Paradigm (NEP)

Table B.8: Effect of NEP on playing Close to Cooperation

Dependent Variable:	Close to Cooperation					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
NEP	0.0279*	-0.0059	0.0283**	0.0038	0.0046**	-0.0002
	(0.0147)	(0.0225)	(0.0123)	(0.0256)	(0.0022)	(0.0037)
T-OnePlayer × NEP		0.0973**		0.0419		0.0099*
		(0.0379)		(0.0400)		(0.0057)
T-LowThreshold × NEP		0.0427		0.0326		0.0065
		(0.0295)		(0.0283)		(0.0046)
Constant	-3.031**	-1.462	-2.806***	-1.503	-0.0770	0.1617
	(1.327)	(1.506)	(1.024)	(1.473)	(0.1914)	(0.2180)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	1,692	1,692	1,692	1,692	1,692	1,692
R^2					0.14517	0.15137
Pseudo R^2	0.09881	0.10619	0.07978	0.08198		
Wald (joint nullity)	5.2641	4.9003	4.8565	4.7357	7.3428	7.5625

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: NEP measures pro-environmental orientation (Dunlap et al., 2000).

Table B.9: Effect of NEP on playing Close to Defection

Dependent Variable:	Close to Defection					
	period 1		period 5		average 1-5	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Logit	Logit	Logit	Logit	OLS	OLS
Variables						
NEP	-0.0441**	-0.0163	-0.0271**	-0.0329*	-0.0018**	-0.0020
	(0.0214)	(0.0327)	(0.0128)	(0.0180)	(0.0008)	(0.0018)
T-OnePlayer × NEP		-0.0055		0.0558		0.0006
		(0.0482)		(0.0603)		(0.0021)
T-LowThreshold × NEP		-0.0581		0.0040		0.0002
		(0.0396)		(0.0250)		(0.0022)
Constant	5.252**	3.807	3.406**	3.587**	0.4819***	0.4887***
	(2.303)	(2.403)	(1.447)	(1.707)	(0.1261)	(0.1522)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	1,692	1,692	1,692	1,692	1,692	1,692
R^2					0.07813	0.07819
Pseudo R^2	0.09485	0.09993	0.08722	0.08848		
Wald (joint nullity)	3.4517	3.0074	5.4242	4.9883	3.2490	2.8520

Clustered (Group) standard errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017): "When should you adjust standard errors for clustering?" Tech. rep., National Bureau of Economic Research.

Anderson, K. and G. Peters (2016): "The trouble with negative emissions," *Science*, 354, 182–183.

Battaglini, M., S. Nunnari, and T. R. Palfrey (2016): "The Dynamic Free Rider Problem: A Laboratory Study," *American Economic Journal: Microeconomics*, 8, 268–308.

Berger, S. and A. M. Wyss (2021): "Measuring pro-environmental behavior using the carbon emission task," *Journal of Environmental Psychology*, 75, 101613.

Brockway, P. E., S. Sorrell, G. Semieniuk, M. K. Heun, and V. Court (2021): "Energy efficiency and economy-wide rebound effects: A review of the evidence and its implications," *Renewable and sustainable energy reviews*, 141, 110781.

Brodeur, A., N. M. Cook, J. S. Hartley, and A. Heyes (2024): "Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement," Tech. rep., I4R Discussion Paper Series.

Cadigan, J., P. T. Wayland, P. Schmitt, and K. Swope (2011): "An experimental dynamic public goods game with carryover," *Journal of Economic Behavior & Organization*, 80, 523–531.

Cadsby, C. B. and E. Maynes (1999): "Voluntary provision of threshold public goods with continuous contributions: experimental evidence," *Journal of Public Economics*, 71, 53–73.

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

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Croson, R. T. and M. B. Marks (2000): "Step returns in threshold public goods: A meta- and experimental analysis," *Experimental Economics*, 2, 239–259.

Dal Bó, P. and G. R. Fréchette (2018): "On the determinants of cooperation in infinitely repeated games: A survey," *Journal of Economic Literature*, 56, 60–114.

Dannenberg, A., A. Löschel, G. Paolacci, C. Reif, and A. Tavoni (2011): "Coordination under threshold uncertainty in a public goods game," *ZEW-centre for european economic research discussion paper*.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2008): "Representative trust and reciprocity: Prevalence and determinants," *Economic Inquiry*, 46, 84–90.

Dunlap, R., K. V. Liere, A. Mertig, and R. E. Jones (2000): "Measuring endorsement of the new ecological paradigm: A revised NEP scale," *Journal of Social Issues*, 56, 425–442.

Fréchette, G. R. and S. Yuksel (2017): "Infinitely repeated games in the laboratory: Four perspectives on discounting and random termination," *Experimental Economics*, 20, 279–308.

Gächter, S., F. Mengel, E. Tsakas, and A. Vostroknutov (2017): "Growth and inequality in public good provision," *Journal of Public Economics*, 150, 1–13.

Gill, D. and V. Prowse (2016): "Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis," *Journal of Political Economy*, 124, 1619–1676.

Hauser, O., D. Rand, A. Peysakhovich, and M. Nowak (2014): "Cooperating with the future," *Nature*, 511, 220–223.

Herr, A., R. Gardner, and J. M. Walker (1997): "An experimental study of time-independent and time-dependent externalities in the commons," *Games and Economic Behavior*, 19, 77–96.

Kimbrough, E. O. and A. Vostroknutov (2015): "The social and ecological determinants of common pool resource sustainability," *Journal of Environmental Economics and Management*, 72, 38–53.

Lackner, K. S., R. Aines, S. Atkins, A. AtKisson, S. Barrett, M. Barteau, R. J. Braun, J. Brouwer, W. Broecker, J. B. Browne, et al. (2016): "The promise of negative emissions," *Science*, 354, 714–714.

Lange, F. (2022): "Behavioral paradigms for studying pro-environmental behavior: A systematic review," *Behavior Research Methods*, 1–23.

Ledyard, J. O. (1995): "Public Goods: A Survey of Experimental Research," in *Handbook of Experimental Economics*, Princeton University Press, 111–94.

Levhari, D. and L. J. Mirman (1980): "The great fish war: an example using a dynamic Cournot-Nash solution," *The Bell Journal of Economics*, 322–334.

Ljungqvist, L. and T. J. Sargent (2018): *Recursive macroeconomic theory*, MIT press.

Maples-Keller, J. L., R. L. Williamson, C. E. Sleep, N. T. Carter, W. K. Campbell, and J. D. Miller (2019): "Using Item Response Theory to Develop a 60-Item Representation of the NEO PI–R Using the International Personality Item Pool: Development of the IPIP–NEO–60," *Journal of Personality Assessment*, 101, 4–15, pMID: 29087223.

Maxwell, S. E., H. D. Delaney, and K. Kelley (2017): *Designing experiments and analyzing data: A model comparison perspective*, Routledge.

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

Ostrom, E. (1990): *Governing the commons: The evolution of institutions for collective action*, Cambridge university press.

Portner, H. O., D. C. Roberts, H. Adams, C. Adler, P. Aldunce, E. Ali, R. A. Begum, R. Betts, R. B. Kerr, R. Biesbroek, et al. (2022): "Climate change 2022: impacts, adaptation and vulnerability," *Genebra, Suíça*.

Proto, E., A. Rustichini, and A. Sofianos (2019): "Intelligence, personality, and gains from cooperation in repeated interactions," *Journal of Political Economy*, 127, 1351–1390.

——— (2022): "Intelligence, errors, and cooperation in repeated interactions," *The Review of Economic Studies*, 89, 2723–2767.

Przepiorka, W. and A. Diekmann (2020): "Binding contracts, non-binding promises and social feedback in the intertemporal common-pool resource game," *Games*, 11, 5.

Rockenbach, B. and I. Wolff (2019): "The Dose Does it: Punishment and Cooperation in Dynamic Public-Good Games," *Review of Behavioral Economics*, 6, 19–37.

Strulik, H. and V. Grossmann (2024): "The economics of aging with infectious and chronic diseases," *Economics & Human Biology*, 52, 101319.

Szekely, A., F. Lipari, A. Antonioni, M. Paolucci, A. Sánchez, L. Tummolini, and G. Andrighetto (2021): "Evidence from a long-term experiment that collective risks change social norms and promote cooperation," *Nature communications*, 12, 1–7.

Tavoni, A., A. Dannenberg, G. Kallis, and A. Löschel (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.

Vespa, E. (2020): "An experimental investigation of cooperation in the dynamic common pool game," *International Economic Review*, 61, 417–440.

Volk, S., C. Thöni, and W. Ruigrok (2011): "Personality, personal values and cooperation preferences in public goods games: A longitudinal study," *Personality and individual Differences*, 50, 810–815.

Walker, J. M. and R. Gardner (1992): "Probabilistic destruction of common-pool resources: experimental evidence," *The Economic Journal*, 102, 1149–1161.

Walker, J. M., R. Gardner, A. Herr, and E. Ostrom (2000): "Collective choice in the commons: Experimental results on proposed allocation rules and votes," *The Economic Journal*, 110, 212–234.