Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2017/23

Experimental Social Planners: Good Natured, but Overly Optimistic

Christoph Engel
Svenja Hippel

MAX PLANCK SOCIETY

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2017/23

Experimental Social Planners: Good Natured, but Overly Optimistic

Christoph Engel / Svenja Hippel

November 2017

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Experimental Social Planners: Good Natured, but Overly Optimistic[*]

Christoph Engel / Svenja Hippel

Abstract

Public goods are dealt with in two literatures that neglect each other. Mechanism design advises a social planner that expects individuals to misrepresent their valuations. Experiments study the provision of the good when preferences might be non-standard. We introduce the problem of the mechanism design literature into a public good experiment. Valuations for the good are heterogeneous. To each group we add a participant with power to impose a contribution scheme. We study four settings: the authority has no personal interest and (1) valuations are common knowledge or (2) active participants may misrepresent their types; the authority has a personal interest (3) and must decide before learning her own valuation or (4) knows her own valuation. Disinterested social planners predominantly choose a payment rule that gives every group member the same final payoff, even if misrepresentation is possible. Authorities are overly optimistic about truth telling. Interested social planners abuse their power, except if the opportunity cost of a more balanced rule is small.

JEL: C91, D02, D03, D61, D62, D64, H23, K12

Keywords: Public Good, Social Planner, Truthtelling, Experiment

[*] Helpful comments by Claudia Cerrone, Daniel Martin, Peter Klibanoff, Joel Sobel and Nan Zhang, and audiences at Northwestern University, the ESA World Meeting 2017 and the 9th NYU CESS Conference are gratefully acknowledged.
Corresponding author: Prof. Dr. Christoph Engel, Max-Planck-Institute for Research on Collective Goods, Kurt-Schumacher-Straße 10, D 53113 Bonn, Germany, engel@coll.mpg.de

1. Introduction

Across disciplinary boundaries, it is not uncommon that literatures do not speak to each other, despite the fact that they deal with closely related issues. Within the same discipline, such mutual neglect is more surprising. This paper deals with one such instance. Public goods are a classic of welfare economics (for a systematic report see Cornes and Sandler 1996). Modern welfare economics has mechanism design foundations. In a mechanism design perspective, the normative problem originates in the heterogeneity of preferences. If a social planner were just to ask individuals how highly they value the provision of a certain public good, she would not learn their true valuations. Individuals would anticipate that the provision of public goods has to be financed from taxes. Anticipating that they will be taxed proportionally to their valuation, statements would systematically be below the true valuation. The first best cannot be attained. Note that this prediction even holds if the mechanism designer can rely on the sovereign powers of the state. In the language of mechanism design, the prediction thus even holds if there is no participation constraint. As long as valuations are private information, even a sovereign mechanism designer must first write an incentive-compatible mechanism that forces individuals to reveal their valuations.

There is also an experimental literature on public goods (for summaries see Ledyard 1995, Zelmer 2003, Chaudhuri 2011). This literature typically assumes away the problem that is central to the mechanism design literature. Typically valuations are induced by the design of the experiment, homogeneous and common knowledge. The literature is interested in the willingness of individuals to disregard the dilemma. It in particular investigates the degree by which alternative types of social preferences support positive contributions. Second generation publications focus on minimal interventions that help communities sustain cooperation, like communication, face to face interaction, or peer punishment.

In this paper we build a bridge between these two literatures. We introduce the problem of the mechanism design literature into the setup of a classic public good experiment. We have four research questions that build on each other. For each question we run a different treatment: How do experimental social planners decide upon the level of providing a public good, and about the allocation of cost

1. if valuations for the good are heterogeneous?
2. if additionally individuals can misrepresent their types ?
3. if their decision affects themselves and they have to decide without knowing their own type?
4. if their decision affects themselves and they know their own type?

From a behavioral perspective, answers are not obvious. The mechanism design literature is interested in efficiency. It thus implicitly defines welfare as the norm. But it stands in conflict with a fair distribution of payoffs, a fair allocation of the cost, and status quo. With the first treatment we learn the empirical distribution of normative preferences in the face of heterogeneous valuations. The information asymmetry defines the mechanism design problem. Yet in a behavioral perspective, this constraint has additional dimensions. Misrepresenting one's valuation requires lying, to which experimental participants have been shown to be averse. And the asymmetry also affects the distributional balance, to which participants in experimental games have shown to pay deference. We study these effects in the second part of the experiment.

Real authorities are usually not completely disinterested. They at the least envisage that, at some later point, the rule that they are now implementing will affect themselves, or their supporters for that matter. To investigate this complication of the mechanism design problem, we repeat the otherwise identical experiment, but have a later participant in the public good game decide without knowing her own type. In other instances, the individual in charge, or her supporters, already know in which ways the rule affects themselves. To elicit these rule choices, we have participants choose a rule before they learn whether they are singled out as “blue” or “red” players, but conditional on them holding “blue” or “red” valuations for the public good. Blue players have a small endowment, but a high marginal per capita rate. Red players have a large endowment, but a low marginal per capita rate.

In the baseline, authorities predominantly choose a rule that levels out heterogeneity and gives every active participant the same final payoff. A small minority of authorities, however, prefers the welfare maximizing solution. Only one of the 32 experimental authorities implements the rule that has every active participant make the same contribution to the public good. Interestingly, many experimental authorities stick to these preferences if active participants get a chance to misrepresent their types. These choices are mainly driven by the beliefs authorities hold about the willingness of active participants to truthfully reveal their types. Authorities strongly overestimate this willingness.

If active participants have to choose a rule without knowing their own type, their choices are split about evenly between efficiency and payoff equality. Again equality of payment is only chosen very rarely. This choice, which would be normative under common knowledge of rationality, becomes a bit more frequent if we maintain the veil of ignorance, but introduce the possibility to misrepresent types. Yet there are still many authorities that choose the rule that is efficient if participants truthfully reveal their types, and many others that choose the rule that equalizes payoff provided active participants do not lie.

In the final step, we let active participants choose a rule conditional on their type. If participants can enforce this rule because lying is excluded by design, the same participants who had made fairly balanced choices when not knowing their own type become straightforwardly selfish, at least if their type is such that selfishness has a high payoff.

To the best of our knowledge, we are the first to experimentally test a social planner in the situation that is prototypical for the mechanism design literature. Traub, Seidl et al. (2009) test a “social planner”, but on pure allocation choices. Rockenbach and Wolff (2016) have a different research question. Observing participants over a whole term, they study which rules participants develop over time. Other papers investigate how risk influences allocation choices of social planners. Cettolin, Riedl et al. (2016) find that uninvolved third parties allocate more to a person who is exposed to a lottery, compared to another who receives the certainty equivalent. Rohde and Rohde (2011) find that social planners prefer allocations where each recipient’s risky allocation is independently drawn, compared to all participants facing one and the same lottery. In Cappelen, Konow et al. (2013) a third party needs to allocate pooled money from two players that made more or less risky choices before. The allocation decisions take place under full information after the lotteries where played out and knowing whether risk takers were lucky or not.

Several papers introduce private information into an experimental public good. Isaac and Walker (1998) do not explicitly inform participants in a linear public good that other group members have the same marginal per capita rate. This does not change results, compared with a baseline where this information is made explicit. Uncertainty about the marginal per capita rate reduces contributions (Levati, Morone et al. 2009), except if the minimum marginal per capita rate still allows for efficiency gains (Levati and Morone 2013). If marginal per capita rates are heterogeneous, this reduces contributions to the public good. Contributions are further reduced if there is uncertainty regarding this heterogeneity (Fischbacher, Schudy et al. 2014). Incomplete information on aggregate contributions to the public good slightly reduces contributions (Chan, Mestelman et al. 1999). If participants have heterogeneous endowments and this affects their profit maximizing choice, contributions to the public good increase. Adding uncertainty about others' endowments does not change these results (Chan, Mestelman et al. 1999). Our experiment differs from all these earlier studies in the dependent variable: we are not interested in the effect of heterogeneity or uncertainty on voluntary contributions, but on rule choice.

Reuben and Riedl (2013) aim at eliciting contribution norms in a heterogeneous public good. Survey respondents strongly favor equality of contributions if the group is homogeneous. If the marginal benefit from the public good is unequal, equality of earnings is the modal choice. If endowments are unequal, contributions proportional to the endowment are most frequently prescribed. These preferences are in line with punishment choices when the mirror games are played out in the lab. Kube, Schaube et al. (2015) show that heterogeneity makes it more difficult for groups to agree on a mechanism that implements the efficient outcome. Our experiment goes beyond in that we have authorities with explicit power to rule and, most importantly, manipulate whether the authority faces agents with the ability to conceal their type.

A small literature tests the reactions of experimental participants to mechanisms for the provision of public goods that would be efficient with common knowledge of rationality. Healy (2006) shows that participants best respond to the last observation they are making in a repeated game. Güth, Koukoumelis et al. (2014) investigate in which ways the reactions of experimental participants to mechanisms depend on them being perceived as fair. Robbett (2016) shows that a mechanism improves contributions to a public good even if there is room for lying. These results help us predict the reactions of participants to interventions by our experimental authorities.

The remainder of this paper is organized as follows: in section 2, we introduce the design of the baseline and a series of post-experimental tests that help us identify motives. We report which contribution rules experimental authorities prefer, and how these choices can be explained. In section 3, we investigate in which ways uninvolved experimental authorities respond to the risk that active participants misrepresent their types. In section 4, we compare the choices of uninvolved authorities with those by involved authorities who decide without knowing their type, both with common knowledge of valuations for the public good, and when these valuations are private information. In section 5, we investigate the rule choices of involved blue and red authorities, again with or without common knowledge of the valuations of others. Section 6 compares choices across treatments. Section 7 concludes with discussion.

2. Uninvolved Unconstrained Authority

The mechanism design literature defines the problem of a social planner as choosing the best rule, given valuations for the public good are heterogeneous and individuals can misrepresent their type. From a behavioral perspective, we must unpack the problem. Before (in the next section) we can study the social planner's reactions to the risk of misrepresentation, we must learn how experimental authorities decide if valuations for the public good are heterogeneous. From a behavioral angle, it is not obvious that they impose the efficient outcome. They might balance out efficiency and fairness concerns. This is what we study in this first treatment.

a) Design

Our baseline is a one-shot linear public good. Hence profit π_i is given by

$$\pi_i = e_i - c_i + \mu_i \sum_{k=1}^{K} c_k \tag{1}$$

where e is the endowment, c is the contribution to the public good, μ is marginal per capita rate, i is the active group member in question, and k is any group member, including the member in question. We implement two-dimensional heterogeneity. There are two blue participants, with $e_b = 100, \mu_b = .6$, and two red participants, with $e_r = 250, \mu_r = .4$. This distribution of valuations is public information. In the baseline group members are passive. To each group, a fifth participant is randomly assigned. In the instructions, this participant is referred to as the "green" player. For her participation in the main experiment, this participant is remunerated with 15 €. Her income does not depend on choices she or other participants make during the experiment. The red and blue participants are informed about this. The green player knows that, for the red and blue participants, either the first or the second part of the experiment is payoff relevant. When making her choice in the first part (the baseline), the green player does not know what the second part is about. The active participants do only know that the green players make choices in the baseline that may affect their payoff, but do not learn any details. All valuations and profits during the experiment are expressed in experimental currency units ECU.

It is the task of the green player to select one of four contribution rules. If the green player chooses *max*, participants of either type must contribute their complete endowments. The second option is *eqinc* (for “equal income”). Red players have to contribute 150, while blue players must contribute 100. The third option is *eqpay*. It obliges participants to make an "equal payment" of 100. The final option *zero* does not oblige active participants to contribute anything to the public project. Figure 1 summarizes the consequences for payoffs. In the interest of making sure that green players understand the implications of their choices, they receive a table with the resulting payoffs for active members.[1]

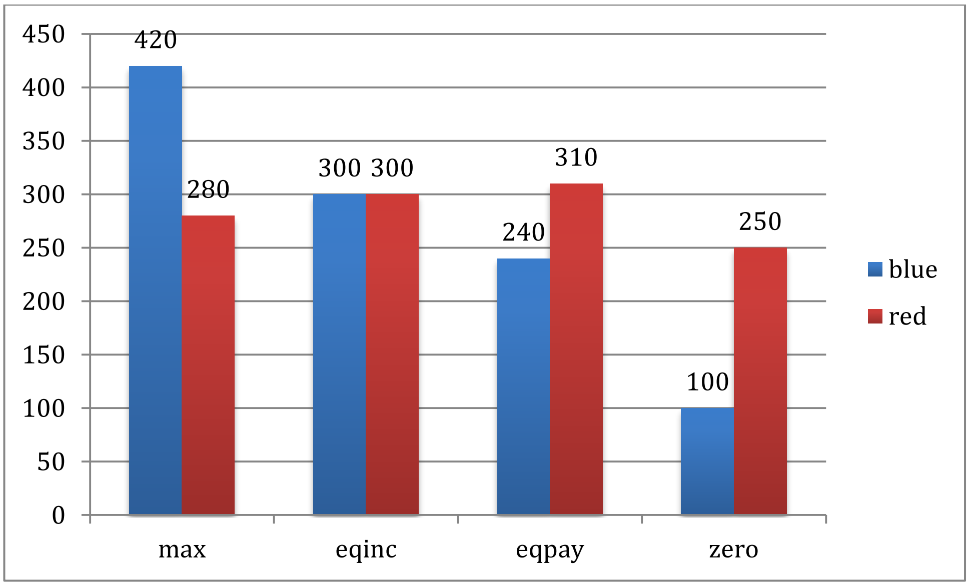

Figure 1
Payoff per Decision Rule

In the interest of having more scope for explaining choices, after the main experiment, we administer three incentivized additional tests.[2] We first have players guess how many active members of their group tell the truth when given a chance to lie, conditional on the chosen rule. If they get this number right, they earn an additional Euro each. To test for lying aversion, we use a procedure developed by Gneezy, Rockenbach et al. (2013).[3] Participants are randomly matched to new pairs of two. Using the strategy method (Selten 1967), participants decide in the role of participant A and of participant B in two separate pairings. After the experiment, roles are randomly assigned. The computer randomly assigns an integer number between 1 and 6 to the pair, but does not inform them at that point. Participant A decides, once more using the strategy method, which signal to send to participant B, conditional on the true number. Participant B decides for any possible signal whether to follow or not. Participant A earns 100 ECU + 20 * the signal. Participant B earns 30 ECU if she does not follow. If she follows, she earns 100 ECU if the signal is correct, and 0 otherwise. We finally elicit social value orientation, using the slider measure developed by Murphy and Ackermann (2014).[4] All feedback is withheld until the end of the entire experiment, to preserve independence.

[1] You can find the table in the instruction displayed in appendix B.

[2] We also use the 10item version of the Big5 inventory (Rammstedt and John 2007), ask trust questions from the German Socio Economic Panel, and request demographic information. We only use the trust questions for data analysis.

[3] We are grateful to Le Quement and Marcin (2016) for sharing their zTree code with us.

[4] We gratefully acknowledge using the code provided and explained by Crosetto, Weisel et al. (2012).

The experiment was run in 2016 in the EconLab of Bonn University. The experiment was computerized, using Fischbacher (2007). Participants were invited using Bock, Baetge et al. (2014). 160 students of various majors participated. They were randomly assigned to 32 groups of five (four active participants and one authority). 56.88% were female, mean age 23.77. Participants on average earned 14.11 € (14.93 $ on the first day of the experiment), 15.97 € for authorities, and 13.64 € for active participants.

b) Hypotheses

The choice of decision rule does not have payoff consequences for the green player. She is free to impose the norm she deems fit. If she is exclusively interested in her own payoff, she is indifferent. She would choose a rule at random. We do however not deem this likely. As one of us has shown in another experiment, disinterested experimental authorities who have power to punish active players in a symmetric linear public good use this power to discipline freeriders (Engel and Zhurakhovska 2017). Likewise, social planners in Cappelen, Konow et al. (2013) and in Cettolin, Riedl et al. (2016) have made meaningful, responsible choices. By analogy we expect social planners in our experiment to aim at making normatively desirable choices. Yet given the asymmetry, it is not obvious though which choice is most desirable. The rules from which the authority must choose are meant to capture prominent competing norms.

The Rule *max* is efficient (for efficiency as a fairness norm see Rabin 1993, Engelmann and Strobel 2004). But this rule is very favorable for blue players and much less favorable for red players. Since they have a smaller endowment, blue players contribute less to the public project. And since they have a higher marginal per capita rate, they benefit more from contributions than red players. Hence if the ruler, in the spirit of inequity aversion, cares about relative payoffs, this rule is less appealing.[5] Note, however, that *max* is at the Pareto frontier. With this rule, red players earn 30 ECU more than their endowment. But blue players earn 320 ECU more than their endowment.

eqinc is mapped onto outcome based definitions of fairness, and inequity aversion in particular (Fehr and Schmidt 1999, Bolton and Ockenfels 2000). If the authority cares about minimizing inequitable outcomes, she will choose this rule. But given the heterogeneity of payoff functions, this outcome can only be achieved if red players contribute 50 ECU more than blue players. Yet this only corresponds to 60% of red players' endowment, whereas blue players are obliged to contribute fully.

In fairness terms, *eqpay* keeps the burden of contributing to the public good constant (cf. Cardenas, Stranlund et al. 2002, Gampfer 2014). Note that this is also the option a mechanism designer would choose in an environment where types are private information.

[5] For a formal definition of the ruler's utility assuming aversion against inequitable outcomes, and the implications for the choice between *max* and *eqinc*, see below section 3 b).

Finally in fairness terms, *zero* conserves the status quo ante (cf. Mandler 2004, Masatlioglu and Ok 2005, Ortoleva 2010): red players are ahead of blue players by a ratio of 5:2.

We have no reason to expect that all experimental authorities adhere to the same fairness interpretation of the situation. The design of the experiment is meant to make each interpretation plausible.

c) Results

The predominant choice was *eqinc* (19), followed by *max* (10), whereas only a single authority chose *eqpay*, and only two authorities chose *zero*. *eqinc* is the choice that is in line with outcome based fairness. Social value orientation is an established measure for this fairness preference (Liebrand and McClintock 1988, Murphy and Ackermann 2014). In the baseline authorities have power to impose a distribution of outcomes. It is therefore remarkable that authorities' personal social value orientation does not explain their choice of rule.[6] This suggests that authorities do not try to impose their personal policy preferences, but rather try to match what they believe is the predominant preference of active participants.[7]

3. Uninvolved Authority Constrained by the Risk of Misrepresentation

The mechanism design literature is interested in reactions of a social planner to her ignorance about individuals' valuation for the public good. This literature assumes that individuals will misrepresent their type if this increases their profit. The literature tries to "design mechanisms" that incentivize individuals to reveal their types. From a behavioral perspective it is neither obvious that individuals will lie about their valuation, nor that social planners will expect them to lie, and impose rules that are not vulnerable to this possibility. This is what we investigate in our second treatment.

a) Design

This treatment differs from the previous by this one element: active participants are asked for their type, and free to report their type as "blue" or "red". This signal determines how much they have to contribute to the public project, depending on the rule chosen by the authority. However the marginal per capita rate is determined by their true type. This is why red players have an incentive to represent their type as "blue" if the authority chooses *max* or *eqinc*. They are indifferent between telling the truth and misrepresenting their type if the authority chooses *eqpay* or *zero*. Blue players do not have an incentive to send a "red" signal. Actually we exclude that they report being "red" if the authority chooses

[6] Coef -.007, p = .830. In all parametric estimations of rule choices, we use ordered logit, and code *max* = 1, *eqinc* = 2, *eqpay* = 3, *zero* = 4. This order reflects (a) the degree of efficiency, (b) the distributional advantage for the blue player, (c) the degree by which the status quo ante is altered, and in treatments with the possibility to make false statements (d) gains from misrepresentation for the red players.

[7] This interpretation is in line with Engel and Zhurakhovska (2016).

max or *eqinc*. In either case they would have to contribute more than their endowment to the public project. Groups of five stay constant over the entire (main) experiment. This is known to participants. For the red and blue players, either the first or the second part of the experiment is payed out, with equal probability.

b) Hypotheses

If authorities want to maintain the status quo, or if they want to make sure that each active participant has an equal share in the provision of the public project, the risk of misrepresentation is immaterial. In the former case, the authority would still choose *zero*, in the latter case she would still choose *eqpay*. This yields

> **Hypothesis 1:** If authorities choose *zero* or *eqpay* when types are common knowledge, they stick to this choice when types are private information.

If authorities exclusively strive for efficiency, the risk of misrepresentation is also immaterial. If both red players lie, *max*, *eqinc* and *eqpay* lead to the same outcomes. If at least one red player tells the truth, total income is highest with *max*. If the choice of *max* when types were common knowledge has been driven by a preference for efficiency, we expect

> **Hypothesis 2**: If authorities choose *max* when types are common knowledge, they stick to this choice when types are private information.

If authorities assume common knowledge of rationality, they lose power to impose their normative convictions. Given that both red players lie and send a “blue” signal, *max* and *eqinc* rules become pointless. The achieved payoffs are identical with *eqpay*. But common knowledge of rationality is a strong assumption. It only holds if all active participants with red valuations are happy to lie. This would run counter evidence showing that a substantial fraction of experimental participants are unwilling to lie, at least if the cost of telling the truth is not prohibitive (Erat 2013, Rauhut 2013, Abeler, Becker et al. 2014, Robbett 2016). Participants with red valuations might also prefer not to lie because they hold social preferences themselves. For either reason, experimental authorities might believe that at least a fraction of red participants will tell the truth. Then the normative assessment becomes more involved. Figure 2 summarizes in which ways payoffs are affected by the choice of rule and the number of red players who tell the truth.

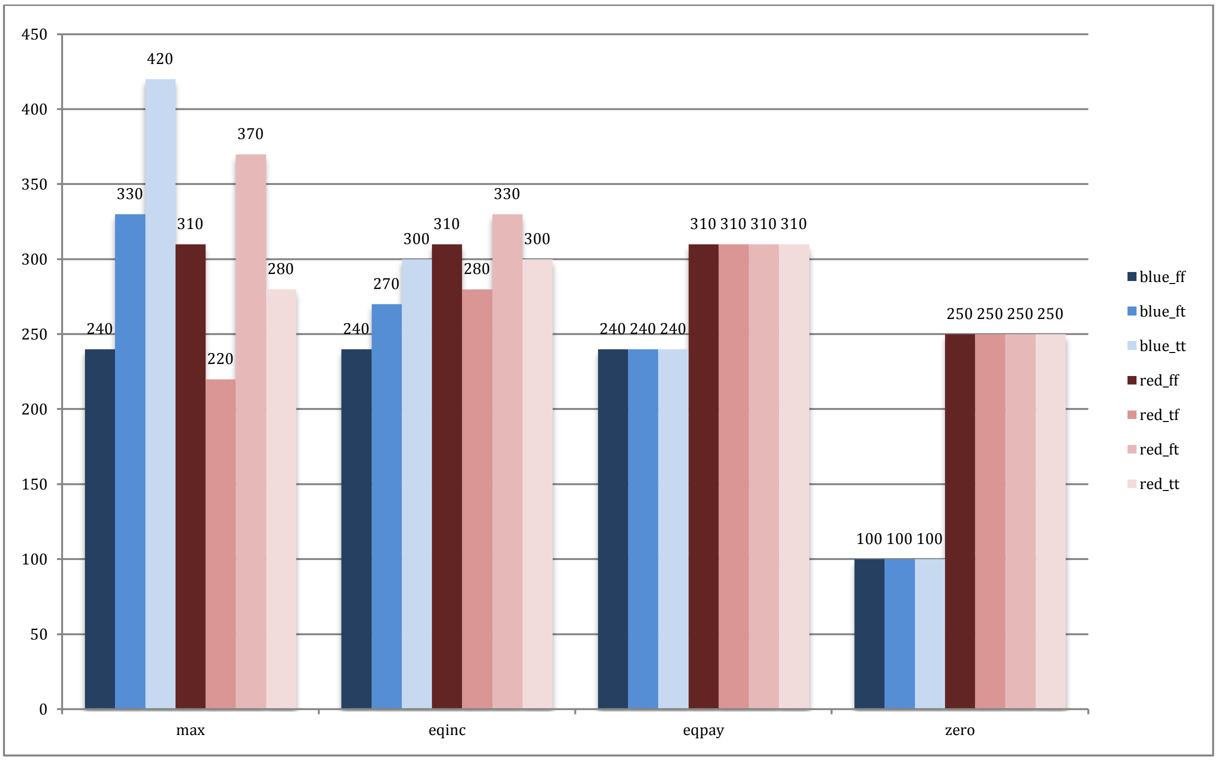

Figure 2
Payoffs Conditional on Chosen Rule and Truth Telling
codes: ff: both red players lie; tt: both tell the truth; tf/ft: one lies (and for a red player, the first is this player's report)

For the assessment of the desirability of *max* or *eqinc* in the face of behavioral uncertainty, the authority has to take more concerns into account. She can rely on the same normative considerations that matter if she has power to impose her choice. But once there is room for lying, intentions come into play. If the rule is *max* or *eqinc* and a red player decides to lie, this player not only violates the moral duty to tell the truth. She also not only deprives the society of welfare, and blue players of a benefit they should have had in the opinion of the authority. The liar additionally exposes the other red player to the risk of deliberate exploitation. The green player has authority to protect a loyal red player, through choosing a rule that makes the risk of misrepresentation immaterial.

The risk of misrepresentation is critical if authorities aim at achieving outcome equality. We now define this preference formally. By the design of the experiment, the authority has no personal pecuniary interest. Her utility must be defined in terms of results she wants to achieve for the four active participants. In the spirit of the model by Fehr and Schmidt (1999) the (purely psychological) utility of the green player u_g with respect to all four active participants of the group is defined by (2)[8]

$$u_g = \sum_{i=1}^{2}(\pi_r^i + \pi_b^i) - \frac{\gamma}{3}\sum_{j=1}^{2}\sum_{i=1}^{2}\left|\pi_r^i - \pi_b^j\right| - \frac{\delta}{3}\left|\pi_r^1 - \pi_r^2\right| \qquad (2)$$

[8] See appendix D for a derivation of (2).

where g stands for the green player, b for a blue player, and r for a red player. π is the actual payoff of an active participant. $\gamma \geq 0$ and $\delta \geq 0$ capture disutility from payoff differences between the players. The authority has to trade off efficiency (the sum of payoffs, captured by the first term) against inequity. The second term represents disutility from payoff differences between the two player types, red and blue. The double sum covers all distances between each of the two (active) red and the two (passive) blue players. The utility function does not distinguish between advantageous and disadvantageous inequity (α and β in the original Fehr/Schmidt model), since at the group level, the inequity is of necessity two-sided, and the green player is not herself affected, so that the direction of the disutility cancels out. Note that the two blue players always get an equally high payoff and therefore the distance $\left|\pi_b^1 - \pi_b^2\right|$ is always 0. Therefore (2) has no additional term to capture inequality among blue players. The second term includes all discrepancies in payoffs that evolve from the respective payment rule itself and additionally disutility from exploiting the passive blue players. Their exploitation risk is limited by the design of the payoff structure though, because the payoff of a blue player under *max* and *eqinc* can never fall below the level that *eqpay* would ensure them. The third term represents disutility from payoff differences between the two red players. The third term only matters when exactly one of the red players misrepresents her type. Arguably $\delta \geq \gamma$: it is normatively more problematic to let down a rule-abiding participant that could have protected herself by lying as well.

Note that, with (2), the authority's utility is the same as welfare if active players hold standard Fehr/Schmidt utility functions (with identical parameters) and $\gamma = \delta = \alpha + \beta$.[9] The four utilities of the active players then sum up to (2) and this can be interpreted as the choice function of a social planner who aims at finding the optimal rule for a society of inequity averse individuals.

Assuming (2), the authority's utility from choosing either rule is as in Table 1. If types are common knowledge, *eqpay* and *zero* are dominated, irrespective of γ. The authority chooses *max* as long as $\gamma < \frac{15}{14}$. She chooses *eqinc* otherwise. If types are unknown and the authority expects one red player to lie, but does not differentiate between inequity to the detriment of a blue and a red player, she chooses *max* as long as $\gamma = \delta < \frac{5}{6}$. This is only slightly more demanding than if she expects both red players to tell the truth. If the authority attaches more weight to an honest red player being let down ($\delta > \gamma$), she prefers *max* over *eqinc* as long as $\delta < \frac{3}{2} - \frac{4}{5}\gamma$. She prefers *eqpay* over *eqinc* if $\delta > \frac{3}{2} + \frac{7}{5}\gamma$. *Zero* remains dominated. Finally if the authority expects both red players to lie, she is indifferent between *max, eqinc* and *eqpay* while *zero* is still dominated.

[9] Taking into account that, by design, the two blue players always have the same payoff, so that inequity in their relationship can be neglected.

	max	$eqinc$	$eqpay$	$zero$
tt	$1400-\frac{560}{3}\gamma$	1200	$1100-\frac{280}{3}\gamma$	$700-200\gamma$
tf	$1250-100\gamma-100\delta$	$1150-\frac{140}{3}\gamma-\frac{100}{3}\delta$	$1100-\frac{280}{3}\gamma$	$700-200\gamma$
ff	$1100-\frac{280}{3}\gamma$	$1100-\frac{280}{3}\gamma$	$1100-\frac{280}{3}\gamma$	$700-200\gamma$

Table 1
Ruler's Utility Assuming Inequity Aversion
tt: both red players tell the truth, tf: one red player tells the truth, ff: both red players pretend to be blue

Based on this definition of utility, assuming that the authority believes that precisely one red player tells the truth and that $\gamma < \frac{15}{14}, \delta > \frac{3}{2} - \frac{4}{5}\gamma$,[10] we expect

> **Hypothesis 3**: If authorities choose *max* when types are common knowledge, they shift to *eqinc* when types are private information.

If $\gamma > \frac{15}{14}$ and $\delta < \frac{3}{2} + \frac{7}{5}\gamma$, we expect

> **Hypothesis 4**: If authorities choose *eqinc* when types are common knowledge, they stick to this choice when types are private information.

For $\gamma > \frac{15}{14}, \delta > \frac{3}{2} + \frac{7}{5}\gamma$, we have the competing

> **Hypothesis 5**: If authorities choose *eqinc* when types are common knowledge, they shift to *eqpay* when types are private information.

c) Results

Figure 3 compares choices of authorities without and with the risk of misrepresentation. We have only three authorities who chose either *eqpay* or *zero* in the first part of the experiment. None of them sticks to her choice. But these are not enough observations to test Hypothesis 1.[11]

[10] Note that this inequality is derived for the case that the authority attaches more weight to an honest red player being let down, hence $\delta > \gamma > 0$ also needs to hold.

[11] Due to a group size of 5, we only have 32 independent observations for the choices of authorities. If a norm is not popular among experimental authorities, we only have few observations. This is why, in this part of the paper, for some norms we can only report descriptive statistics. By the design of the experiment, for the choices of interested authorities, we have four times as much data, and can engage in more fine-grained analysis.

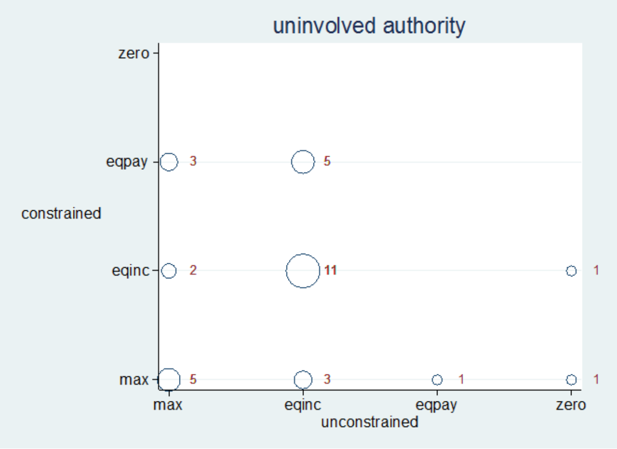

Figure 3
Rule Choices of Uninvolved Authorities Without and With the Risk of Misrepresentation
bubble size indicates frequency

When there was no risk of misrepresentation, 10 authorities chose *max*. Of the five authorities who stick to this choice in the face of possible misrepresentation of type, three believe that there will be no misrepresentation, and one thinks that only one red player will report a "blue" type. This is in line with Hypothesis 2. But we also have one player who expected both red players to report truthfully and nonetheless shifts to *eqpay*. For a statistical test of the hypothesis, we do not have enough observations.[12]

We do not have support for Hypothesis 3: from the 10 authorities that had chosen *max* when they could impose this choice on all active participants, only two switch to *eqinc* when active participants may misreport their types.

When there was no risk of misrepresentation, 19 authorities chose *eqinc*. 11 stick to this choice. Nine of them deem it certain that there is no misrepresentation. This is in line with Hypothesis 4. Of the 6 authorities who believe that only one of the red players reports truthfully, 4 shift to *eqpay*. This is in line with Hypothesis 5. We also find statistical support for this result. If an authority believes that only one red player reports truthfully, she is 59.82% more likely to shift to *eqpay*, and 38.26% less likely to stick to her earlier choice.[13]

We conclude

> **Result 1:** Uninvolved authorities with a preference for equalizing outcomes only shift to equal payments if they deem it likely that at best one of two active participants will tell the truth.

[12] Note that we have much more statistical power for choices of involved neutral authorities, see sections 4 and 5 below.

[13] Marginal effects from an ordered logit regression. The regression explains choices of authorities when there is a risk of misrepresentation (second treatment of the experiment) and who had chosen *eqinc* in the first treatment of the experiment. The explanatory variable is the authority's belief about the number of truth telling red participants, coef 3.264, p = .015. 6 of these 19 authorities believe that only one red player tells the truth. The remaining 13 authorities believe that both red players report their types truthfully.

Table 2 summarizes the beliefs of authorities, split by their choice in the first treatment of the experiment, and possible choices in the second treatment. Authorities are fairly optimistic. More than a third (11/32) believe that no active player will lie, even if the rule is *max*. Two thirds (21/32) believe that all active players will tell the truth if the rule is *eqinc*. Authorities who have chosen *eqinc* in the first treatment are most optimistic. None of these 19 authorities believes that both red players will lie if they again chose *eqinc*.

rule in treatment 1	*max*		*eqinc*		*eqpay*		*zero*	
# truth tellers	*max*	*eqinc*	*max*	*eqinc*	*max*	*eqinc*	*max*	*eqinc*
0	4	2	7	0	0	0	1	0
1	2	3	6	6	1	0	0	0
2	4	5	6	13	0	1	1	2

Table 2
Beliefs of Uninvolved Authorities, By Unconstrained and Potential Constrained Choice

Actually, authorities are heavily overoptimistic. Active participants had to indicate their choice, conditional on each of the four possible rules, before their type was determined, which is why we have 128 observations for each rule. 110 of these participants decided to misreport their type if the rule would be *max*, and 101 decided to misreport their type if the rule would be *eqinc*. This gives us

> **Result 2:** Uninvolved authorities overestimate the willingness of active participants to tell the truth.

The regressions in Table 3 show what drives active participants' decision whether to tell the truth.[14] Their own social preferences (their social value orientation scores) do not explain these choices. If they are more prone to lying on the post-experimental test for lying aversion, they are a bit less likely to tell the truth about their type.[15] Their willingness to trust the statements by their random counterpart in the post-experimental test for lying aversion does not explain truth telling in the main experiment.[16] By contrast, we find a strong positive effect of beliefs. The more participants deem it likely that others in their group will tell the truth, the more they are willing to tell the truth themselves. This suggests what creates the mismatch between authorities' beliefs and actual truth telling: authorities do not seem to sufficiently factor in defensive lying: red participants make a false statement for fear of being let down by the other red player.

[14] Choices to lie if the rule is *max* or if it is *eqinc* come from the same participant. This source of dependence is captured by a multivariate regression that allows residual errors from both choices per participant to be correlated. Since the dependent variables are binary, we estimate logit coefficients.

[15] The raw data from this test consist of 6 choices per participant. We compress them into a single measure by running a regression, separately for each participant, that explains the number she reports on that test with the signal she receives. The coefficients from these local regressions do not turn out informative, while the constants do. The constant is small if the participant reports a small number if the signal is small, i.e. if the participant tells the truth.

[16] This measure is generated in an equivalent way. We run a (linear) local regression that explains, separately for each participant, the willingness to follow the statement with the signal. We again use the constant from this local regression. Using the coefficient of the signal does not turn out informative either.

	max				*eqinc*			
	model 1	model 2	model 3	model 4	model 1	model 2	model 3	model 4
SVO score	.014 (.018)	.004 (.018)	.008 (.019)	.016 (.020)	.013 (.015)	.002 (.016)	.007 (.016)	.019 (.017)
active lying		-.181^{+} (.106)	-.201^{+} (.110)	-.207^{+} (.116)		-.196* (.090)	-.228* (.095)	-.203^{+} (.105)
trust			-.469 (.456)	-.345 (.495)			-.654^{+} (.397)	-.700 (.445)
belief				2.117** (.711)				2.595** (.782)
cons	-2.035*** (.390)	-1.444** (.483)	-1.131* (.568)	-2.564** (.818)	-1.517*** (.327)	-.869* (.415)	-.430 (.495)	-2.527** (.840)
N	128	128	128	128	128	128	128	128

Table 3
Explaining Truth Telling to Uninvolved Authority
multivariate logit
dvs: dummy that is 1 if participant decides to reveal that her type is red, conditional on either rule being in place
SVO score: angle from slider measure; active lying: constant of local regression, explaining statement with signal; trust: constant of local regression, explaining following the statement with signal; belief: how many active group members with red valuations will tell the truth, given the respective rule
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

We conclude

> **Result 3:** Participants lie more about their type the more they believe that other participants will lie.

4. Involved Authority Deciding Under the Veil of Ignorance

Many constitutions constrain the legislator to the adoption of general rules. Its choices should not be ad hoc. Ideally, statutes are abstract, and strike a balance for a multiplicity of conflicts of life. This is why rule making authorities often do not know in which way the rule they are adopting today may affect themselves (or their partisans) at some future point in time. But they know that they are not outside the law. The rule may therefore affect themselves as well. In this treatment, we investigate in which way experimental social planners react to this uncertainty.

a) Design

The green player no longer participates in the third and fourth part of the experiment. In the third part of the experiment, ex post one of the four active players is randomly singled out as authority, with equal probability. Yet she has to choose a rule without knowing her own type.

Participants know that a fourth part of the experiment is to follow, and that only one of the two parts will be paid out, with equal probability, but they do not yet know what the fourth part will be about. In this fourth part, authorities still decide under the veil of ignorance. But then active players have the opportunity to misreport their type. The authority may then also misreport her own type. When choosing what to report, participants do not yet know which group member will be singled out as authority.

b) Hypotheses

In this part of the experiment, rule choices affect the profit of the authority herself. This double role adds another level of complication, even when participants cannot lie. An involved authority might need to trade off the general motive she prefers as an authority with her own profit. The authority chooses the rule under the veil of ignorance. She knows that, with 50% probability, she will have either red or blue valuations. Therefore the authority must think of her utility as if she has either role in the group and it is still helpful to utilize (2) whereby the own profit motive might shift the weights between the terms.

If the authority exclusively cares about her own profit, the expected payoff in the third part of the experiment is 350 if she chooses *max*, 300 if she chooses *eqinc*, 275 if she chooses *eqpay*, and 175 if she chooses *zero*. We predict

> **Hypothesis 6**: When involved authorities decide under the veil of ignorance and types are common knowledge, they choose *max*.

If authorities care about efficiency, there is no tradeoff. They should choose *max* a fortiori. If they prefer an equal sharing of the burden, or if they want to preserve the status quo, they must balance out these motives with the profit motive. The more pronounced these competing motives, the more they are likely to prefer *eqpay* or *zero* over *max.* The same holds for *eqinc*, if they care about equality of outcomes.

If participants now have the possibility to misreport their types, Hypothesis 1 still holds. If the authority prefers *eqpay* or *zero*, there is no reason to deviate from this choice when types are private information. Also if authorities expect all red players to lie about their type then there is still no scope for *max* or *eqinc*.

For uninvolved authorities, the possibility that red players lie reduces the scope for enforcing their normative convictions. This is why, in section 3, we have focused the analysis on the effect of beliefs. Involved authorities face an additional concern. If it turns out that their own type is red and they have chosen *max* or *eqinc*, they must decide whether to tell the truth themselves. Authorities that care more about their own payoffs may see this again as an opportunity to maximize profit. They can also shield themselves actively from potential exploitation through the other red player lying. If this is the dominant effect we should see

> **Hypothesis 7:** When involved authorities decide under the veil of ignorance and types are private information, the authority is more likely to choose *max* than when types are common knowledge.

By contrast authorities might be hesitant to violate their own rule, be that *max* or *eqinc*, but also loathe having reduced their own income, both in absolute and in relative terms. If this effect is dominant, we should see

> **Hypothesis 8:** When involved authorities decide under the veil of ignorance and types are private information, the authority is less likely to choose *max* and *eqinc* than when types are common knowledge.

c) Results

Under the veil of ignorance and in the absence of the risk of misrepresentation, 62 participants choose *max*, 61 choose *eqinc*, while only 4 choose *eqpay*, and only a single participant chooses *zero*. This result is in line with Hypothesis 6: almost half of the choices maximize expected payoff. These choices could, however, also be driven by a preference for efficient outcomes. If we explain rule choices with the individual's social value orientation score, we do not find a significant effect. We do however find that uninvolved authorities are 15.04% more likely to choose *eqinc* than involved authorities deciding under a veil of ignorance.[17] We interpret this as tentative evidence in support of Hypothesis 6. We conclude

> **Result 4**: Involved authorities are more likely to choose *max* than uninvolved authorities if types are common knowledge and they do not yet know their own valuation of the public good.

As Figure 4 shows, more than half of all involved authorities do not change the rule if participants have a chance to lie: 79 of 128 choices are on the diagonal. This is in line with Hypothesis 2 and 4. There are again too few observations for *eqpay* and *zero* to test Hypothesis 1. 9 authorities switch from *max* to *eqpay*, and 10 authorities switch from *eqinc* to *eqpay*. These 19 choices are in line with Hypothesis 5. 8 of the former and 8 of the latter group of authorities believe that at most one red player will tell the truth, were they to maintain their earlier choice.[18]

[17] Average marginal effect from an ordered logit regression, explaining rule choices with the fact that the authority is uninvolved, N = 160, coef of authority being uninvolved .773 p = .054, marginal effect p = .045.

[18] Note that these are beliefs about choices when the authority is uninvolved, though.

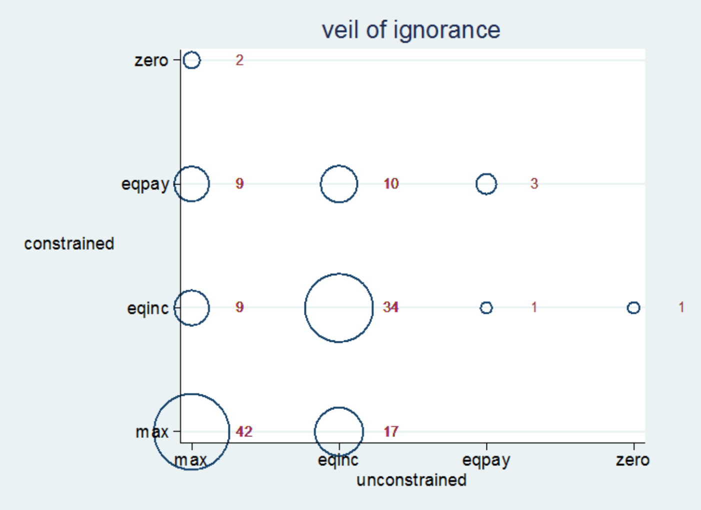

Figure 4
Rule Choices of Involved Authorities Deciding Under the Veil of Ignorance, Without and With the Risk of Misrepresentation
bubble size indicates frequency

Overall, the regressions in Table 4 are better in line with Hypothesis 8 than with Hypothesis 7. The regressions predict that the average authority shifts away from *max* when there is room for lying. The coefficient "constrained" is significant at conventional levels if we control for the willingness of the authority to lie herself if the rule is *eqinc*, and interact it with the presence of a lying opportunity (model 2).[19] Average marginal effects from these regressions are even more revealing. If there is room for lying, overall authorities are 9.37% less likely to choose *max* (model 1, p = .044), 5.13% more likely to choose *eqinc* (p = .051), and 3.63% more likely to choose *eqpay* (p = .058). With model 2, we can separately calculate average marginal effects for authorities that later tell the truth and those that lie. Those that lie are 11.48% less likely to choose *max* (p = .025), 6.00% more likely to choose *eqinc* (p = .036), and 4.70% more likely to choose *eqpay* (p = .034).[20] Controlling for beliefs does not yield additional insights (model 3). Involved authorities show little sensitivity towards the risk that others might lie about their valuations.

[19] Results look similar if we replace the willingness to lie if the rule is *eqinc* by the willingness to lie if the rule is *max*.

[20] Average marginal effects for authorities who later tell the truth are all insignificant.

	model 1	model 2	model 3
constrained	.530^{+} (.272)	.646* (.295)	.646* (.295)
lie if *eqinc*		-.473 (.642)	-.448 (.662)
constrained * lie if *eqinc*		-.823 (.774)	-.824 (.774)
expected fraction of red players telling the truth if *eqinc*			-.070 (.450)
cut 1	.130 (.238)	.047 (.255)	.020 (.309)
cut 2	3.061*** (.388)	2.993*** (.395)	.2965*** (.430)
cut 3	5.709*** (.722)	5.678*** (.726)	5.620*** (.745)
N	256	256	256

Table 4
Explaining Rule Choice by Involved Authorities Deciding Under the Veil of Ignorance
random effects ordered logit
dv: rule, coded 1 max, 2 eqinc, 3 eqpay, 4 zero
constrained: participants may misrepresent their valuation
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

Yet in Figure 4 we also have 17 authorities that had chosen *eqinc* when lying was ruled out by design, but choose *max* once lying becomes possible. 14 of them later decide to lie if the rule is *max* and they have red valuations. This is precisely the logic behind Hypothesis 7. We therefore conclude

> **Result 5**: The majority of uninvolved authorities do not react to the fact that red players can lie about their valuation. A larger minority shift towards a less efficient rule. A smaller minority shift towards *max*.[21]

5. Involved Authority Knowing Her Valuation

Even if the constitution wants the legislator to act for the public benefit, putting aside personal interests of the members of Parliament, this ideal is not necessarily attained in reality. Not so rarely, those in power have partisan interests, and take them into account when choosing which public goods to provide, and how to distribute the cost. In the final treatment, we investigate this situation experimentally.

[21] Since we were afraid that participants would have a hard time spelling out beliefs for four different situations (decision by green authority, by red or blue authority under the veil of ignorance, by a red or a blue authority knowing their own valuation), we have only elicited beliefs for the first situation. We can therefore not investigate whether involved authorities are as purely calibrated as uninvolved authorities.

a) Design

In the fifth and sixth part of the experiment, and before giving participants feedback about the earlier parts of the experiment, we have each participant choose a rule for the entire group assuming that she herself has either blue or red valuation. We have participants take these decisions first assuming that participants cannot misrepresent their type, and then allowing for lying. Afterwards, one participant is randomly singled out. The choices of this participant in the role of authority are implemented.

b) Hypotheses

The involved authority still, as in section 4, needs to trade off the general motive she prefers as an authority with her own payoff. But participants now have more information as they already know their type before deciding about the rule. Therefore in (2) the authority can identify which payoffs and payoff differences directly affect herself. It is no longer justified to just collapse the standard disutility parameters α and β from Fehr and Schmidt (1999) into the symmetric parameters γ and δ. In fact it is easiest to think about (2) as the (weighted) sum of four classical Fehr-Schmidt preferences (one for each group member including the authority). We add weight parameters to the authorities own payoffs and payoff differences that directly affect her. When she is concerned about the general motive she wants to implement in her role as authority, these weights are balanced out and all four terms count the same. On the other hand, when she is solely interested in her own payoffs, the weight on other payoffs not directly belonging to her is 0. Note that this does not imply that the authority is straightforwardly selfish; only that she may hold classical Fehr-Schmidt preferences looking at payoff differences solely from her own point of view instead of considering the summed up version.

If a participant only considers her own preferences and has blue valuation, her profit is highest (420) if the rule is *max*. If they were to choose *eqinc* instead, they could avoid advantageous inequity. Yet rationalizing *eqinc* instead of *max* would require $\beta > 9/7$.[22] Empirically this is implausibly high (Blanco, Engelmann et al. 2011). The concern for an equal sharing of the burden would have to be even more pronounced for an authority with blue valuation to prefer *eqpay* over *max*. And likewise an authority concerned about maintaining status differences could only prefer *zero* over *max* with an even higher inequity aversion. We therefore have a clear prediction:

> **Hypothesis 9**: When involved authorities know their own type to be blue and types are common knowledge, they choose *max*.

The equivalent exercise for players with red valuation yields a less clear prediction. The profit maximizing choice is *eqpay*. It gives red players a payoff of 310 ECU. But the payoff difference between *eqpay* and the next profitable choice *eqinc* is only 10, while *eqpay* leads to advantageous inequity of 70 ECU. Participants with $\beta > 3/14$ would prefer *eqinc*. Even the payoff difference between *eqpay* and *max* is not huge. At a price of 30 ECU for themselves, red authorities can achieve efficiency. Since the efficiency gain is substantial, this only requires

[22] Solve $420 - \frac{2}{3}\beta(420 - 280) = 300$ for β.

$\alpha > 9/14$ combined with an even smaller β. By contrast, it is highly implausible that a red authority chooses *zero*. We therefore have competing predictions:

> **Hypothesis 10**: When involved authorities know their own type to be red and types are common knowledge,
>
> a) authorities choose *eqpay,*
> b) authorities choose *eqinc,*
> c) authorities choose *max*.

If lying is possible, Hypothesis 1 still holds and when authorities assume common knowledge of rationality, there is still no scope for *max* or *eqinc*. Yet if they believe that at least some red players tell the truth, and their own type is blue, they still stand a chance to maximize their own profit and achieve efficiency by choosing *max*. They can even feel all the more comfortable with this choice as profit maximization does not require that they violate their own rule. We therefore predict

> **Hypothesis 11**: When involved authorities know their own type to be blue, they stick to their choice when types are private information.

Again the situation is more complicated for red authorities. If they are happy to violate their own rule, they should shift to *max*. This gives them the highest payoff. Yet shifting to *eqpay* pays a triple dividend: the choice is incentive compatible for all group members; they do not have to violate their own rule; conditional on them being unwilling to violate the rule this choice gives them the highest payoff. We therefore have competing predictions

> **Hypothesis 12**: If authorities know their own type to be red and participants can misrepresent their types,

a) authorities choose *max,*
b) authorities choose *eqpay.*

c) Results

As the left panel of Figure 5 shows, we straightforwardly support Hypothesis 9: if authorities know their own type to be blue and lying is excluded, 100 of 128 authorities choose *max*. The right panel demonstrates that the choices of authorities knowing their own type to be red are heterogeneous. About half of them (61 of 128) choose *eqpay* and thereby maximize their own profit. 49 choose *eqinc* instead and pay the small price of 10 for greater equality. However only 16 accept the still small price of 30 to achieve efficiency (but also accept disadvantageous inequity). This gives us support for the first two competing statements in Hypothesis 10. Apparently each of them captures a relevant fraction of the population. We conclude

> **Result 6**: When participants cannot lie about their types and
>
> a) the authority's type is blue, she chooses *max,*
> b) the authority's type is red, she either chooses *eqpay* or *eqinc*.

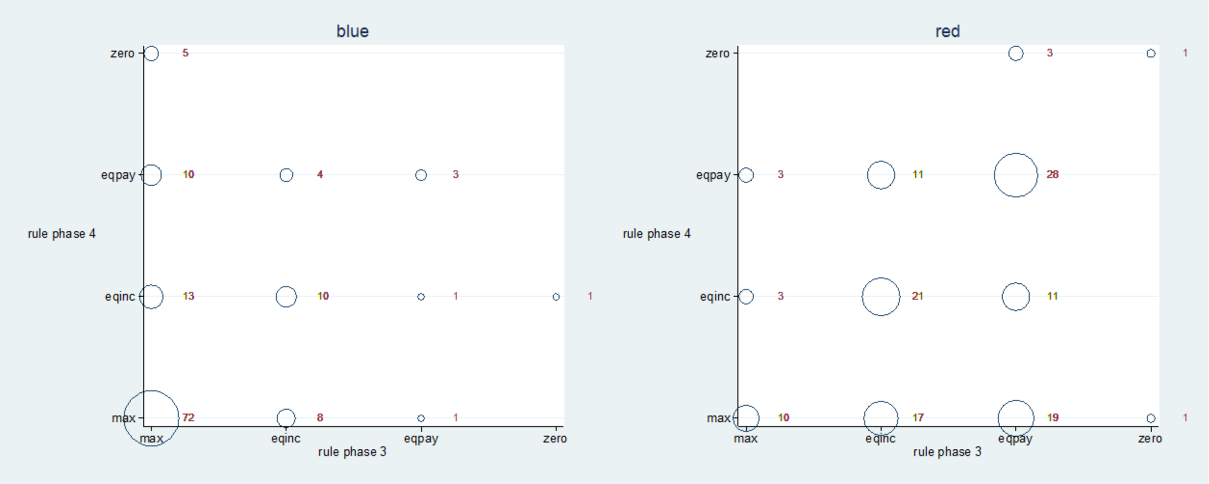

Figure 5
Rule Choices of Authorities Knowing Their Own Type,
Without and With the Risk of Misrepresentation
bubble size indicates frequency

Even if participants may misrepresent their types, the majority of blue authorities (72 of 128) still choose *max*. This choice maximizes their expected profit as long as they do not believe that all red participants will lie. Indeed 45 of the 81 authorities that make this choice believe that at least one of the red participants reveals her type.[23] This supports Hypothesis 11. Much like the green players, the blue authorities are overly optimistic, though. Actually if the rule is *max*, on the condition that their own valuation is blue, only 17 of 128 participants reveal their types.

Almost half of the red authorities (59 of 128) do not change the rule if misrepresentation is possible, including 28 authorities who had chosen *eqpay* when participants could not disguise their type. 11 authorities shift from *eqinc* to *eqpay*, 3 from *max* to *eqpay*. Hence about a third of the authorities behave in line with Hypothesis 12b. Yet even more authorities (47 of 128) either stick to their earlier choice of *max* or, even more importantly, shift to this choice when it becomes possible to disguise one's type. This behavior is in line with Hypothesis 12a. We must have a mixed population and conclude

Result 7: When participants have the possibility to report the wrong type

a) and authorities have the blue valuation themselves, they predominantly choose *max,*
b) and authorities have the red valuation themselves, the majority of them make the same choice as when participants cannot misreport; sizeable minorities shift to either *max* or *eqpay*.

[23] Note however that we have elicited these beliefs for active players' choices when the rule is defined by the uninvolved green player.

6. Comparisons

In section 3, we investigate how the choices of uninvolved authorities react to the risk that active group members might misrepresent their type. In section 4 we do the same for authorities who are involved, but have to decide before they learn their own valuation for the public good. In section 5 we do this for authorities who know their valuation right from the start. In section 3, we also investigate what explains active members' choice to lie about their valuation, if the design allows for this. In this section we add comparisons across treatments, first for authorities, and then for active group members.

Rule choices of uninvolved authorities and of involved authorities deciding under a veil of ignorance are very similar. At conventional significance levels, we only find that involved authorities are 8% less likely than uninvolved authorities to choose *eqinc* when deciding under a veil of ignorance.[24] Authorities that also decide on behalf of themselves strongly react to knowing their own valuation (model 1 of Table 5). The action space of authorities is not continuous. But it arguably is ordered. From *max* to *eqinc* to *eqpay* to *zero* rules become less and less efficient. If misrepresentation is ruled out by design, in the same order rules also become less and less favorable for the blue type. This is why, in Table 5, we estimate ordered logit models. Since we have six (four) choices per authority, we estimate random effects models. The coefficients of ordered logit models are genuinely hard to interpret. This is why, in the following text, we report average marginal effects.

Model 1 predicts that authorities who know their valuation to be red and do not face the constraint are 37% less likely to choose *max*, 29% more likely to choose *eqpay*, and 5% more likely to choose *zero*. If they know their valuation to be blue and do not face the constraint, they are 24% more likely to choose *max*, 14% less likely to choose *eqinc*, 9% less likely to choose *eqpay*, and 1% less likely to choose *zero*. This is statistical support for the fact that authorities that are themselves affected by the chosen rule are very sensitive to its effect on their own payoff. If we pool the data over types (model 1), the regression predicts that the risk of misrepresentation shifts choices away from the individually most profitable choice.

[24] p = .048; this additional regression is available from the authors upon request.

	model 1	model 2
red	2.059*** (.252)	
blue	-1.301*** (.288)	-2.758*** (.716)
constrained	.399 (.244)	-1.464* (.616)
red*constrained	-1.315*** (.347)	
blue*constrained	.678^{+} (.387)	2.864** (.967)
zero type		.165 (.479)
positive type		.389 (.553)
blue*zero type		-.373 (.791)
blue*positive type		.040 (.880)
constrained*zero type		.437 (.679)
constrained*positive type		1.630* (.789)
blue*constrained*zero type		-1.475 (1.080)
blue*constrained*positive type		-1.086 (1.207)
cut 1	.209 (.195)	-1.396** (.439)
cut 2	2.133*** (.218)	.167 (.430)
cut 3	5.174*** (.352)	3.261*** (.517)
N	768	512

Table 5
Explaining Rule Choices Across Treatments
random effects ordered logit
dv: rule, coded 1 max, 2 eqinc, 3 eqpay, 4 zero
constrained: participants may misrepresent their valuation
zero type: when deciding under a veil of ignorance, participant did not react to risk of misrepresentation
positive type: when deciding under a veil of ignorance, participant has reacted to risk of misrepresentation by shifting towards a less efficient rule
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

We get a more differentiated picture if, in model 2, we use the decisions of authorities when they do not know their valuation to predict how they will decide when they have this information at the moment of choosing a rule. Specifically we classify an authority as a “zero type” if this authority, under a veil of ignorance, does not react to the possibility that participants might misrepresent their valuation. We classify an authority as a “negative type” if she shifts towards a more efficient rule, and as a “positive type” if she shifts towards a less efficient rule. We again report average marginal effects.

Under the veil of ignorance, shifting towards *max* is profitable if the authority is happy to lie (and thereby to violate her own rule) should her valuation turn out to be red. This selfish motive is also visible in the choices these authorities make when they know their valuation right from the start. If this valuation is red, they are 30% more likely to choose *max*, 25% less likely to choose *eqpay*, and 3% less likely to choose *zero*. When they have blue valuation themselves, they are 29% less likely to choose *max*, 16% more likely to choose *eqinc*, and 13% more likely to choose *eqpay*.[25]

Under the veil of ignorance, shifting towards a less efficient rule shows that the authority is sensitive towards the possibility that other participants might conceal their valuations. If they learn that their own valuation is red, their rule choices are not significantly affected. Yet if they learn that their own valuation is blue, they are 43% less likely to choose *max*, 14% more likely to choose *eqinc*, 26% more likely to choose *eqpay*, and 3% more likely to choose *zero*. This is further support for a type of authorities that do not consider violating their own rule, but that are concerned about non-authorities being prepared to do so. Note that this risk is most relevant if the authority has blue valuation herself (and therefore nothing to gain from misrepresentation), but now faces two group members who might engage in lying.

Finally we learn more about those authorities who did not adjust the rule to the risk of misrepresentation when deciding under a veil of ignorance. These authorities do not shift either when having blue valuation. But when they have red valuation, they are 19% more likely to choose *max*, 19% less likely to choose *eqpay*, and 3% more likely to choose *zero*. This pattern fits a type that is sensitive to the chance for abusing power, but only falls for it if this is sufficiently profitable.

We finally compare the decision of active participants to tell the truth across treatments. If we pool the choices of active players to reveal their valuation across all treatments, we only find that truth telling is substantially and significantly more likely if the rule is either *eqpay* or *zero*. This is not surprising, given that revelation does not have any payoff consequences. Whether the authority also decides on her own behalf does not have an effect.[26] The regression in Table 6 shows that lying is not only defensive when the authority is uninvolved, but that this is also the predominant motive if the authority is involved: the more others a participant believes to tell the truth, the more she is happy to do so herself. If we calculate average marginal effects, we find that active players are 19% more likely to tell the truth if the rule is *max* and they expect one more group member to tell the truth, whether or not the authority decides on her own behalf. We find that active participants are even 33% more likely to tell the truth if the rule is

[25] All these average marginal effects are significant at the 5% level or lower.

[26] This additional regression is available from the authors upon request.

eqinc and the authority is uninvolved. Here we see an effect of authority involvement. But even if the authority also decides on her own behalf, the fact that they believe one more active participant to tell the truth increases their willingness to tell the truth themselves by 23%.[27]

eqinc	-.250 (.956)
eqpay	.561 (1.011)
zero	.722 (1.051)
involved	-.171 (.871)
*eqinc**involved	.129 (1.374)
*eqpay**involved	2.764* (1.289)
*zero**involved	1.820 (1.343)
belief	2.137** (.784)
*eqinc**belief	.129 (1.374)
*eqpay**belief	2.764* (1.289)
*zero**belief	1.820 (1.343)
involved*belief	.118 (1.092)
*eqinc**involved*belief	-.688 (1.633)
*eqpay**involved*belief	-2.950^{+} (1.514)
*zero**involved*belief	-1.841 (1.550)
cons	-3.580*** (.630)
N	1024

Table 6
Explaining Truth Telling
random effects logit
dv: dummy that is 1 if participant decides to reveal that her type is red
involved: the authority is an active player
belief: how many other group members does this player believe to tell the truth, given the rule
standard errors in parenthesis
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

[27] All reported average marginal effects are significant.

7. Conclusion

In welfare economics theory the social planner is just a thought experiment: what would the first best solution look like? If this godlike individual is omniscient, the question is moot. The social planner would perfectly implement the norm imposed on her by the theorist. Traditionally this goal is maximization of welfare. For the theorist, the problem only gets interesting if the social planner is not omniscient. If the social problem consists of deciding on the optimal level of providing a public good, and if one additionally assumes the population to be heterogeneous, attaining first best becomes challenging. Mechanism design theory defines the conditions under which the mechanism designer need not content herself with second best.

The insights produced by welfare economics rest on the assumption that social problems originate in the self-interest of individuals. A large experimental literature has shown that, in many contexts, this assumption is at variance with observed behavior. The standard object of observation in this literature is an individual interacting in an environment with no explicit rules, and in particular with no rule-making authority. This paper is an early attempt at bringing these two literatures into contact with each other.

A first message is rather sobering. The same participants that, in many public goods experiments, have shown to overcome the dilemma themselves and cooperate to a remarkable degree behave very selfishly in this experiment. More than 80% of them pretend to have a small endowment if their endowment is actually large. Our data suggest that this is chiefly defensive lying. Participants are afraid that others will lie, and do not want to be the sucker. We must leave it for future work to test the robustness of this explanation. Alternative explanations include: the cost of telling the truth was too high; voluntary contributions to public goods are less likely in the first place if the population is heterogeneous; individuals are more prepared to be selfish in their dealings with an authority, rather than directly with their peers.

A second finding is more comforting. When neutral authorities face a population that could thwart their attempts at achieving normatively desirable goals, they do not swerve. They also do not simply try to impose their personal predilections; the authority's social value orientation score does not explain choices. Rather they estimate the probability that members of the population might lie. If a normatively more appealing rule can still be implemented, given these estimates, they choose it. Yet unfortunately these authorities are massively overoptimistic. In the experiment we have however withheld all feedback until the very end. Authorities did therefore not have a chance to learn that active participants are quite willing to misrepresent their type if this helps them avoid having to pay for the public good. An obvious follow-up to our experiment would be a repeated game. It is quite plausible that authorities would adjust the choice of rule – which, unfortunately, would however also make the chosen rule normatively less appealing.

Public choice theory rightly points to differences between real rule makers and the social planners of welfare theory. Politicians may simply be corrupt and maximize personal income. Or they may have an independent political agenda that differs from the wishes of the electorate. Two variants of the experiment investigate these qualifications by making the authority an interested agent. Specifically in the first variant the authority knows that she has one of two valuations for the public good herself, but has to decide under the veil of ignorance. This is analogous to a situation where those in power do not yet know in which way they will be affected by some predictable change in circumstances. In this context, a selfish, risk neutral authority maximizes her expected profit. We find only weak support for this hypothesis. The choice that maximizes expected profit is significantly, but only mildly more frequent if the authority also decides on her own future income. If, however, authorities already know their own valuation, we see a very different picture. Those with high valuation for the public good predominantly force all to contribute their entire endowments. Interestingly those with low valuation for the public good exhibit more differentiated choices. Many of them also impose the rule that maximizes their own profit. Yet a sizeable minority decides in favor of socially more desirable rules. The likely explanation for this difference is cost. Individuals with high valuation for the public good lose a lot of profit if they favor equity over efficiency. By contrast those with low valuation for the public good only lose a little bit of personal income if they privilege equality of income over equality of burden. Of course proving this explanation to be true would require a new experiment that manipulates cost.

In the final step, we combine the fact that the authority is herself interested in the outcome with the fact that she must choose a rule without knowing the individual valuation of each addressee. If these authorities decide under the veil of ignorance, only a rather small minority shift to the rule that maximizes personal profit provided they misrepresent their own type. From a normative perspective this is welcome news. At least if authorities are ad hoc, they are hesitant to violate their rules. The majority of these authorities either stick to their preferred choice, or they shift to the rule that also a mechanism designer would have to choose and impose the highest contribution to the public good that is still in line with the preferences of individuals having low valuation for the good.

We note further limitations. Rule choices are not incentivized in the first two parts of the experiment. We have done so for two reasons: the third and fourth parts of the experiment are one way of contrasting unincentivized with incentivized choices. And earlier experiments have shown that experimental authorities make a reasonable attempt at achieving normatively desirable goals (Engel and Zhurakhovska 2016). We have not counterbalanced order. We again have two justifications: we have withheld feedback until the very end of the experiment. Therefore contamination can at most stem from having seen the choice in a different light. Moreover with four different stages, this would have required 24 different treatments. With five participants per independent observation, this would have been unmanageable. Finally many rule choices are not simply noisy about a general mean. Rather effects are heterogeneous. We can discern different groups of authorities following a different logic. Arguably these groups reflect that experimental authorities have come to the lab with different types of personality. This poses two statistical challenges. At least in the ad hoc situation of an experiment, personality cannot be induced, but only measured. We can therefore not rely on random assignment for identification. Moreover we frequently cannot prove, in a strict sense that authorities are of a certain type. We then have to content ourselves with suggestive evidence.

While we acknowledge these limitations, we believe that our experiment is a valuable first step at better understanding the choices of real authorities tasked with securing the provision of a public good when valuations are heterogeneous, as well as the reactions of those under their spell. That way we hope to have somewhat reduced the still large gap between the mechanism design literature on public goods and the experimental literature on public goods.

References

ABELER, JOHANNES, ANKE BECKER AND ARMIN FALK (2014). "Representative Evidence on Lying Costs." Journal of Public Economics **113**: 96-104.

BLANCO, MARIANA, DIRK ENGELMANN AND HANS-THEO NORMANN (2011). "A Within-Subject Analysis of Other-Regarding Preferences." Games and Economic Behavior **72**: 321-338.

BOCK, OLAF, INGMAR BAETGE AND ANDREAS NICKLISCH (2014). "hroot: Hamburg Registration and Organization Online Tool." European Economic Review **71**: 117-120.

BOLTON, GARY E. AND AXEL OCKENFELS (2000). "ERC: A Theory of Equity, Reciprocity and Competition." American Economic Review **90**: 166-193.

CAPPELEN, ALEXANDER W., JAMES KONOW, ERIK Ø. SØRENSEN AND BERTIL TUNGODDEN (2013). "Just Luck. An Experimental Study of Risk Taking and Fairness." American Economic Review **103**: 1398-1413.

CARDENAS, JUAN CAMILO, JOHN STRANLUND AND CLEVE WILLIS (2002). "Economic Inequality and Burden-sharing in the Provision of Local Environmental Quality." Ecological Economics **40**(3): 379-395.

CETTOLIN, ELENA, ARNO RIEDL AND GIANG TRAN (2016). Giving in the Face of Risk.

CHAN, KENNETH S, STUART MESTELMAN, ROBERT MOIR AND R ANDREW MULLER (1999). "Heterogeneity and the Voluntary Provision of Public Goods." Experimental Economics **2**(1): 5-30.

CHAUDHURI, ANANISH (2011). "Sustaining Cooperation in Laboratory Public Goods Experiments. A Selective Survey of the Literature." Experimental Economics **14**: 47-83.

CORNES, RICHARD AND TODD SANDLER (1996). The Theory of Externalities, Public Goods and Club Goods. Cambridge, Cambridge University Press.

CROSETTO, PAOLO, ORI WEISEL AND FABIAN WINTER (2012). A Flexible z-Tree Implementation of the Social Value Orientation Slider Measure (Murphy et al. 2011)–Manual.

ENGEL, CHRISTOPH AND LILIA ZHURAKHOVSKA (2016). You are in Charge. Experimentally Testing the Motivating Power of Holding a (Judicial) Office.

ENGEL, CHRISTOPH AND LILIA ZHURAKHOVSKA (2017). "You Are In Charge. Experimentally Testing the Motivating Power of Holding a Judicial Office." Journal of Legal Studies **46**: 1-50.

ENGELMANN, DIRK AND MARTIN STROBEL (2004). "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments." American Economic Review **94**: 857-869.

ERAT, SANJIV (2013). "Avoiding Lying. The Case of Delegated Deception." Journal of Economic Behavior & Organization **93**: 273-278.

FEHR, ERNST AND KLAUS M. SCHMIDT (1999). "A Theory of Fairness, Competition, and Cooperation." Quarterly Journal of Economics **114**: 817-868.

FISCHBACHER, URS (2007). "z-Tree. Zurich Toolbox for Ready-made Economic Experiments." Experimental Economics **10**: 171-178.

FISCHBACHER, URS, SIMEON SCHUDY AND SABRINA TEYSSIER (2014). "Heterogeneous Reactions to Heterogeneity in Returns from Public Goods." Social Choice and Welfare **43**(1): 195-217.

GAMPFER, ROBERT (2014). "Do Individuals Care about Fairness in Burden Sharing for Climate Change Mitigation? Evidence from a Lab Experiment." Climatic Change **124**(1-2): 65-77.

GNEEZY, URI, BETTINA ROCKENBACH AND MARTA SERRA-GARCIA (2013). "Measuring Lying Aversion." Journal of Economic Behavior & Organization **93**: 293-300.

GÜTH, WERNER, ANASTASIOS KOUKOUMELIS, M VITTORIA LEVATI AND MATTEO PLONER (2014). "Providing Revenue-generating Projects under a Fair Mechanism. An Experimental Analysis." Journal of Economic Behavior & Organization **108**: 410-419.

HEALY, PAUL J (2006). "Learning Dynamics for Mechanism Design. An Experimental Comparison of Public Goods Mechanisms." Journal of Economic Theory **129**(1): 114-149.

ISAAC, R MARK AND JAMES M WALKER (1998). "Nash as an Organizing Principle in the Voluntary Provision of Public Goods. Experimental evidence." Experimental Economics **1**(3): 191-206.

KUBE, SEBASTIAN, SEBASTIAN SCHAUBE, HANNAH SCHILDBERG-HÖRISCH AND ELINA KHACHATRYAN (2015). "Institution Formation and Cooperation with Heterogeneous Agents." European Economic Review **78**: 248-268.

LE QUEMENT, MARK T AND ISABEL MARCIN (2016). Communication and Voting in Heterogeneous Committees. An Experimental Study.

LEDYARD, JOHN O. (1995). Public Goods. A Survey of Experimental Research. The Handbook of Experimental Economics. J. H. Kagel and A. E. Roth. Princeton, NJ, Princeton University Press**:** 111-194.

LEVATI, M VITTORIA AND ANDREA MORONE (2013). "Voluntary Contributions with Risky and Uncertain Marginal Returns. The Importance of the Parameter Values." Journal of Public Economic Theory **15**(5): 736-744.

LEVATI, M VITTORIA, ANDREA MORONE AND ANNAMARIA FIORE (2009). "Voluntary Contributions with Imperfect Information. An Experimental Study." Public Choice **138**(1-2): 199-216.

LIEBRAND, WIM B. AND CHARLES G. MCCLINTOCK (1988). "The Ring Measure of Social Values. A Computerized Procedure for Assessing Individual Differences in Information Processing and Social Value Orientation." European Journal of Personality **2**: 217-230.

MANDLER, MICHAEL (2004). "Status quo Maintenance Reconsidered. Changing or Incomplete Preferences?" Economic Journal **114**(499): F518-F535.

MASATLIOGLU, YUSUFCAN AND EFE A OK (2005). "Rational Choice with Status Quo Bias." Journal of Economic Theory **121**(1): 1-29.

MURPHY, RYAN O. AND KURT A. ACKERMANN (2014). "Social Value Orientation: Theoretical and Measurement Issues in the Study of Social Preferences." Personality and Social Psychology Review **18**: 13-41.

ORTOLEVA, PIETRO (2010). "Status Quo Bias, Multiple Priors and Uncertainty Aversion." Games and Economic Behavior **69**(2): 411-424.

RABIN, MATTHEW (1993). "Incorporating Fairness into Game Theory and Economics." American Economic Review **83**: 1281-1302.

RAMMSTEDT, BEATRICE AND OLIVER P. JOHN (2007). "Measuring Personality in One Minute or Less. A 10-item Short Version of the Big Five Inventory in English and German." Journal of Research in Personality **41**: 203-212.

RAUHUT, HEIKO (2013). "Beliefs about Lying and Spreading of Dishonesty:.Undetected Lies and their Constructive and Destructive Social Dynamics in Dice Experiments." PloS One **8**(11): e77878.

REUBEN, ERNESTO AND ARNO RIEDL (2013). "Enforcement of Contribution Norms in Public Good Games with Heterogeneous Populations." Games and Economic Behavior **77**: 122-137.

ROBBETT, ANDREA (2016). Just Ask. Preference Revelation and Lying in a Public Goods Experiment.

ROCKENBACH, BETTINA AND IRENAEUS WOLFF (2016). "Designing Institutions for Social Dilemmas." German Economic Review **17**: 316-336.

ROHDE, INGRID MT AND KIRSTEN IM ROHDE (2011). "Risk Attitudes in a Social Context." Journal of Risk and Uncertainty **43**(3): 205-225.

SELTEN, REINHARD (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. Beiträge zur experimentellen Wirtschaftsforschung. E. Sauermann. Tübingen, Mohr**:** 136-168.

TRAUB, STEFAN, CHRISTIAN SEIDL AND ULRICH SCHMIDT (2009). "An Experimental Study on Individual Choice, Social Welfare, and Social Preferences." European Economic Review **53**(4): 385-400.

ZELMER, JENNIFER (2003). "Linear Public Goods. A Meta-Analysis." Experimental Economics **6**: 299-310.

Appendix

A. Instructions for Red and Blue Players

Welcome to our experiment!

You can earn a substantial sum of money in this experiment. It is therefore very important that you read the following instructions carefully.

No communication with the other participants is allowed during the experiment. Should you disobey this rule, you will be excluded from the experiment and all payments. If you have any questions, please raise your hand. We will then come to you.

You will be taking part in **4 experiments** today. The individual experiment procedures will be explained to you on your computer screens before they begin.

There are three different roles in all experiments, referred to henceforth as red player, blue player and green player.

The computer has randomly determined that you will always be, in all experiments, **either a red or a blue player**, but never a green player. You will be informed during the experiments which role colour you have been assigned.

During all experiments, we shall speak not of Euro, but of Taler. Your entire income will hence initially be calculated in Taler. The total number of Taler accumulated by you during the experiments will then be converted into euro at the end, at a rate of **1 Euro = 50 Taler**.
The Taler you will have earned during the experiments will be paid out to you in Euro and in cash at the end.

At the end of all experiments, **either the first or the second experiment** will be randomly selected by the computer, and your individual income from it paid out to you. In addition, **either the third or the fourth experiment** will be randomly selected by the computer and also paid out to you.

Payment to all other red and blue players will be carried out in the same way. Payment to the green player, on the other hand, does not depend on the experiments that have been drawn. The green player will be paid a certain sum in Euro in any case, and independently of his or her decisions.

Please note: During all computer selections during the experiments, all available options shall be **equally probable**. We will explain in detail whenever the computer makes a random draw.

First Experiment

You are **not an active player** in this experiment.

Another participant makes a payoff-relevant decision for you in this experiment. You may receive a certain number of Taler as a result of this decision. You will be told once all experiments have ended which decision the other participant has made.
Please click **"OK"** to get to the next experiment.

Second Experiment

For this experiment, the computer has randomly assigned you the role of **(red/blue) player**.
In addition, the computer has randomly assigned you to a **group** of four players in total, made up of **exactly two red and two blue players**. Apart from you, there are therefore three further players in your group, one further (red/blue) player and two (blue/red) players.

Further, the computer has assigned to your group of four a green player, who has also been selected at random.

Each **red player** receives **250 Taler** and each **blue player** receives **100 Taler**, which we will henceforth refer to as an **endowment**. This endowment may be used in various ways. Either it is contributed to a joint project, or it can be retained.

The total income (in Taler) for red and blue players is divided into two parts: (1) the Taler income from the joint project and (2) the retained Taler.

Total income (in Taler) = Income from the joint project + Taler retained

The income from the joint project is calculated from the total sum of all contributions to the project (within the group of four). For the red players, this total sum of all contributions to the project is multiplied by 0.4; for the blue players, it is multiplied by 0.6.

Red player:
Income from the joint project = Total sum of all contributions to the project (within the group of four) x 0.4

Blue player:
Income from the joint project = Total sum of all contributions to the project (within the group of four) x 0.6

If, **for example**, the sum of the contributions of all group members to the joint project is 600 Taler, each red player in the group will receive an income from the project of 0.4 x 600 = 240 Taler. Each blue player in the group will receive an income from the project of 0.6 x 600 = 360 Taler.

If, **for example**, you contribute 100 Taler out of your endowment to your group's project, the sum of all contributions to the joint project rises by 100 Taler and your income from the project, as the income of a red player, rises by 40

Taler or, in the case of a blue player, by 60 Taler. However, this also means that the income of every other group member also rises either by 40 Taler for red players or 60 Taler for blue players, so that the group's total income rises by 2 x 40 + 2 x 60 = 200 Taler.
The other members in your group therefore also profit from your contributions to the project. In turn, you profit from the other group members' contributions to the project. For every Taler that another group member contributes to the project, you earn 0.4 Taler as a red and 0.6 Taler as a blue player.

If, **for example**, each member of your group of four contributes 100 Taler to the project, each red player will receive 4 x 100 x 0.4 = 160 Taler as income from the project, while each blue player will receive 4 x 100 x 0.6 = 240 Taler as income from the project.

In this experiment, the **green player** decides for your group of four which of the **4 payment rules** will be used for the joint project:

Name of payment rule	Red contribution	Blue contribution
Maximum payment	250	100
Equal income	150	100
Equal payment	100	100
Zero payment	0	0

Your **decisions** now always consist of naming a colour. You decide **separately for each of the four payment rules** which colour you wish to name. Each time you can opt for "blue" or for "red". The colour you name **does not have to be the same as the colour assigned to you by the** computer. The other three members of your group have the same task.

The green player finds out only at the end of all experiments how many players in the group of four opted for which colour. Moreover, neither the green player nor the other players in the group find out, either during or after the experiment, which role colour you were assigned by the computer.

Please bear the following in mind: The number of Taler both you and the players in your group of four contribute to the joint project in this experiment does not depend on the colour you actually have, but rather on the colour you name.

If, **for example**, the green player has opted for the payment rule Maximum Payment and all four players have named the role colour "blue", then 100 Taler are automatically taken from the endowment of each player in the group of four as a contribution to the group's joint project. For this is the contribution a blue player makes when the rule Maximum Payment is in force. These decisions hence lead to the sum of the contributions of all group members to the joint project adding up to 400 Taler.

However, the income from the joint project is still calculated on the basis of the player's actual role colour. For a red player, this total sum of contributions to a project is multiplied by 0.4, despite this player having named the colour "blue"; for blue players, it is multiplied by 0.6.

Please note that for the payment rules you can only name colours for which your endowment is still sufficient. As a blue player, for example, you would not be permitted to name the colour "red" for the payment rule Maximum Payment, since your endowment would not be enough to pay a contribution of 250 Taler towards the joint project.

You may use the information leaflets **"support calculations for the second experiment"** for all your decisions. These leaflets should be on the table in your booth.

If you have any questions, please raise your hand at any time. We will come to you.

Third Experiment

Once again you are in the **same group of four** with the same group members as in the second experiment. The green player is inactive in this experiment.

The roles within the group of four are **redistributed** by the computer for this experiment. A player who was a red player in the second experiment may now either be a red or a blue player in the third experiment, regardless of the role this player previously had. In total, however, the group once again consists of two red and two blue players. You will only be told at the end of all experiments which role colour you had in this experiment.

For this experiment, each player in the group of four receives a **new endowment**. Each **red player** once again receives **250 Taler** and each **blue player** once again receives **100 Taler**. Only this new endowment may be used in this experiment, i.e., it can either be paid into a joint project or retained.

In the following, you will make **three decisions for your entire group of four.** The other players are only told how you decided after all experiments have ended.

If the computer randomly selects this experiment for payment at the end, one player from your group of four will be drawn and one of the three decisions made by this player will become payoff-relevant for the entire group. With the same probability, the computer will choose either the first decision (which you will make on your next screen), or else either one of the other two decisions (which you will make on your next two screens). This means your decision can determine the number of Taler which you and the other players in your group of four will receive. You will be given the number of Taler corresponding to your role colour.
You can opt for one of the **4 payment rules** that will be used for your group's joint project:

Name of payment rule	Red contribution	Blue contribution	Total income red	Total income blue
Maximum payment	250	100	280	420
Equal income	150	100	300	300
Equal payment	100	100	310	240
Zero payment	0	0	250	100

If, **for example**, you opt for the payment rule Equal Income, then 150 Taler are automatically taken from the endowment of each red player in the group of four as a contribution to the group's joint project, and 100 Taler are taken from each blue player. Your decision hence leads to the sum of the contributions of all group members to the joint project adding up to 500 Taler.

If you have any questions, please raise your hand at any time. We will come to you.

Fourth Experiment

Once again you are in the **same group of four** with the same group members as in the second and third experiment. The green player is once again inactive.

The roles within the group of four are once again **redistributed** by the computer for this experiment. You will only be told in the second part which role colour you have in this experiment.

For this experiment, each player in the group of four once again receives a **new endowment**. Each **red player** once again receives **250 Taler** and each **blue player** once again receives **100 Taler.** Only this new endowment may be used in this round, i.e., it can either be paid into a joint project or retained.

If the computer randomly selects this experiment for payment at the end, one player from your group of four will be drawn and one of the three decisions made by this player in the **first part** will become payoff-relevant for the entire group. As in the third experiment, the computer will choose, with the same probability, either the first decision (which you will make on the first decision screen), or else either one of the other two decisions (which you will make on your next two screens). Together with the decisions of all group members from the **second part**, this decision will become payoff-relevant for the entire group. This means your decision can determine the number of Taler which you and the other players in your group of four will receive. You will be given the number of Taler corresponding to your role colour.

First Part

In the following, you will make **three decisions for your entire group of four.** The other players are only told how you decided after all experiments have ended.

You can opt for one of the **4 payment rules** that will be used for your group's joint project:

Name of payment rule	Red contribution	Blue contribution
Maximum payment	250	100
Equal income	150	100
Equal payment	100	100
Zero payment	0	0

In the second part of this experiment, you and the other players in your group will once again have the opportunity to **name one colour** for each of the four payment rules – either "blue" or "red". The colour you name **does not have to be the same as the colour assigned to you by the computer**. Just as in the second experiment, the contributions to the joint project are calculated on the basis of the named colour.

However, the income from the joint project is once again calculated on the basis of the player's actual role colour. For a red player, this total sum of all contributions to a project is therefore once again multiplied by 0.4; for blue players, it is multiplied by 0.6.

The randomly chosen payment rule of the randomly chosen player in your group of four will therefore determine, along with the named colours from the second part of this experiment, how many Taler you and the other players will receive for this experiment.

Once again, you may use the information leaflets **"support calculations for the second experiment"** for all your decisions.

If you have any questions, please raise your hand at any time. We will come to you.

Second Part

For this experiment, the computer has randomly assigned you the role of **(red/blue) player**.
Your **decisions** now always consist of naming a colour. You decide **separately for each of the four payment rules** which colour you wish to name. Each time you can opt for "blue" or for "red". The colour you name **does not have to be the same as the colour assigned to you by the computer**. The other three members of your group have the same task.

The other players in your group of four will never be told, neither during nor after the experiment, which role colour you were assigned by the computer.

Please note that for the different payment rules you can once again only name colours for which your endowment is still sufficient.

The randomly chosen payment rule of the randomly chosen player in your group of four from the first part will therefore become payoff-relevant for your entire group, along with the named colours from this part of the experiment. Please note therefore: It may be that one of your own decisions from the first part is drawn, or else a decision made by one of the other players from your group of four.

Once again, you may use the information leaflets **"support calculations for the second experiment"** for all your decisions.

If you have any questions, please raise your hand at any time. We will come to you.

B. Instructions for Green Players

Welcome to our experiment!

You can earn a substantial sum of money in this experiment. It is therefore very important that you read the following instructions carefully.
No communication with the other participants is allowed during the experiment. Should you disobey this rule, you will be excluded from the experiment and all payments. If you have any questions, please raise your hand. We will then come to you.

You will be taking part in **2 experiments** today. The individual experiment procedures will be explained to you on your computer screens before they begin.

There are three different roles in all experiments, referred to henceforth as red player, blue player and green player.

The computer has randomly assigned you the role of **green player**. Your role will remain unchanged during all experiments.

You will receive a total of **15 Euro** for taking part in our experiments today. You will definitely receive this sum in its entirety, independently of the decisions you make during the experiments.

During all experiments, we shall speak not of Euro, but of Taler. The income of the red and blue players will hence initially be calculated in Taler and then converted into euro at the end, at a rate of **1 Euro = 50 Taler**.

At the end of all experiments, **either the first or the second experiment** will be randomly selected by the computer, and the income resulting from it paid out to the red and blue players. Your own payment is not affected by this.

Please note: During all computer selections during the experiments, all available options shall be **equally probable**. We will explain in detail whenever the computer makes a random draw.

First Experiment

You have been randomly assigned by the computer to a **group** of four other players. This group of four consists of **exactly two red and two blue players**.

Each **red player** receives **250 Taler** and each **blue player** receives **100 Taler**, which we will henceforth refer to as an **endowment**. This endowment may be used in various ways. Either it is contributed to a joint project, or it can be retained.

The total income (in Taler) for red and blue players is divided into two parts: (1) the Taler income from the joint project and (2) the retained Taler.

Total income (in Taler) = Income from the joint project + Taler retained

The income from the joint project is calculated from the total sum of all contributions to the project (within the group of four). For the red players, this total sum of all contributions to the project is multiplied by 0.4; for the blue players, it is multiplied by 0.6.

Red player:
Income from the joint project = Total sum of all contributions to the project (within the group of four) x 0.4

Blue player:
Income from the joint project = Total sum of all contributions to the project (within the group of four) x 0.6

If, **for example**, the sum of the contributions of all group members to the joint project is 600 Taler, each red player in the group will receive an income from the project of 0.4 x 600 = 240 Taler. Each blue player in the group will receive an income from the project of 0.6 x 600 = 360 Taler.

If one group member contributes 1 Taler of his or her endowment, the sum of contributions to the joint project rises by 1 Taler. The income from the project rises by 0.4 Taler for each red player and by 0.6 Taler for each blue player. The group's total income therefore increases by 2 x 0.4 + 2 x 0.6 = 2 Taler. All group members therefore profit from each individual group member's contributions to the project.

If, **for example**, each member of your group of four contributes 100 Taler to the project, each red player will receive 4 x 100 x 0.4 = 160 Taler as income from the project, while each blue player will receive 4 x 100 x 0.6 = 240 Taler as income from the project.

In this experiment, however, the red and blue players do not decide themselves. Rather, **you decide for the group of four that has been assigned to you**. The players in the group are only told how you decided after all experiments have ended.

In the following, you can opt for one of **4 payment rules** for the group's joint project:

Name of payment rule	Red contribution	Blue contribution	Total income red	Total income blue
Maximum payment	250	100	280	420
Equal income	150	100	300	300
Equal payment	100	100	310	240
Zero payment	0	0	250	100

If, **for example**, you opt for the payment rule Equal Income, then 150 Taler are automatically taken from the endowment of each red player in the group of four as a contribution to the group's joint project, and 100 Taler are taken from each blue player. Your decision hence leads to the sum of the contributions of all group members to the joint project adding up to 500 Taler.

If you have any questions, please raise your hand at any time. We will come to you.

Second Experiment

Once again you are in the **same group of four** with the same group members as in the first experiment.

The roles within the group of four are **redistributed** by the computer for this experiment. A player who was a red player in the first experiment may now either be a red or a blue player in the second experiment, regardless of the role this player previously had. In total, however, the group once again consists of **exactly two red and two blue players**.

For this experiment, each player in the group of four receives a **new endowment**. Each **red player** once again receives **250 Taler** and each **blue player** once again receives **100 Taler**. Only this new endowment may be used in this experiment, i.e., it can either be paid into a joint project or retained.

In this experiment, the red and blue players once again do not decide themselves, but rather **you decide for the group of four that has been assigned to you**. The players in the group are only told how you decided after all experiments have ended.

In the following, you can opt for one of **4 payment rules** for the group's joint project:

Name of payment rule	Red contribution	Blue contribution
Maximum payment	250	100
Equal income	150	100
Equal payment	100	100
Zero payment	0	0

The red and blue players have already made some decisions in this experiment. They have decided **separately for each of the four payment rules** which colour they wish to name – either "blue" or "red". **The colour that has been named does not have to be the same as the colour assigned to them by the computer**. Only once all experiments have ended are you told which colours the players in the group of four have named.

For example, it is possible that three players have chosen "blue" and one player has chosen "red". It is also possible that all players in the group of four have chosen the colour "red".
Please bear the following in mind: The number of Taler the players in the group of four contribute to the joint project in this experiment does not depend on the colour they actually have, but rather on the colour they name.

If, **for example**, you opt for the payment rule Maximum Payment and all four players have named the role colour "blue", then 100 Taler are automatically taken from the endowment of each player in the group of four as a contribution to the group's joint project. For this is the contribution a blue player makes when the rule Maximum Payment is in force. Your decision hence leads to the sum of the contributions of all group members to the joint project adding up to 400 Taler.

However, the income from the joint project is still calculated on the basis of the player's actual role colour. For a red player, this total sum of contributions to a project is multiplied by 0.4, despite this player having named the colour "blue"; for blue players, it is multiplied by 0.6.

Please note that, for the different payment rules, the players in the group of four can only name colours for which their endowment is still sufficient. A blue player, for example, would not be permitted to name the colour "red" for the payment rule Maximum Payment, since this player's endowment would not be enough to pay a contribution of 250 Taler towards the joint project.

You may use the information leaflets **"support calculations for the second experiment"** for your decision. These leaflets should be on the table in your booth.

If you have any questions, please raise your hand at any time. We will come to you.

C. Support Calculations for the Second Experiment

a) Blue Player

Note: Your own total income is printed in **bold**.
Assume you are a blue player. Assume further that the other members of your group of four name under the following payment rule:

Maximum payment

Number of colours named	**Total income when you name red:**				**Total income when you name blue:**			
	red player named red	red player named blue	blue player named blue	**blue player named red**	red player named red	red player named blue	**blue player named blue**	blue player named red
3 red	-	-	-	-	-	-	-	-
2 red and 1 blue	-	-	-	-	280	-	**420**	-
1 red and 2 blue	-	-	-	-	220	370	**330**	-
3 blue	-	-	-	-	-	310	**240**	-

Equal income

Number of colours named	**Total income when you name red:**				**Total income when you name blue:**			
	red player named red	red player named blue	blue player named blue	**blue player named red**	red player named red	red player named blue	**blue player named blue**	blue player named red
3 red	-	-	-	-	-	-	-	-
2 red and 1 blue	-	-	-	-	300	-	**300**	-
1 red and 2 blue	-	-	-	-	280	330	**270**	-
3 blue	-	-	-	-	-	310	**240**	-

Equal payment

Number of colours named	**Total income when you name red:**				**Total income when you name blue:**			
	red player named red	red player named blue	blue player named blue	**blue player named red**	red player named red	red player named blue	**blue player named blue**	blue player named red
3 red	310	-	-	**240**	310	-	**240**	240
2 red and 1 blue	310	310	240	**240**	310	310	**240**	240
1 red and 2 blue	310	310	240	**240**	310	310	**240**	240
3 blue	-	310	240	**240**	-	310	**240**	-

Zero payment

Number of colours named	**Total income when you name red:**				**Total income when you name blue:**			
	red player named red	red player named blue	blue player named blue	**blue player named red**	red player named red	red player named blue	**blue player named blue**	blue player named red
3 red	250	-	-	**100**	250	-	**100**	100
2 red and 1 blue	250	250	100	**100**	250	250	**100**	100
1 red and 2 blue	250	250	100	**100**	250	250	**100**	100
3 blue	250	-	-	**100**	250	-	**100**	100

b) Red Player

Note: Your own total income is printed in **bold**.
Assume you are a red player. Assume further that the other members of your group of four name under the following payment rule:

Maximum payment

Number of colours named	Total income when you name red:				Total income when you name blue:			
	red player named red	red player named blue	blue player named blue	blue player named red	red player named red	**red player named blue**	blue player named blue	blue player named red
3 red	**-**	-	-	-	-	**-**	-	-
2 red and 1 blue	**-**	-	-	-	-	**-**	-	-
1 red and 2 blue	**280**	-	420	-	220	**370**	330	-
3 blue	**220**	370	330	-	-	**310**	240	-

Equal income

Number of colours named	Total income when you name red:				Total income when you name blue:			
	red player named red	red player named blue	blue player named blue	blue player named red	red player named red	**red player named blue**	blue player named blue	blue player named red
3 red	**-**	-	-	-	-	**-**	-	-
2 red and 1 blue	**-**	-	-	-	-	**-**	-	-
1 red and 2 blue	**300**	-	300	-	280	**330**	270	-
3 blue	**280**	330	270	-	-	**310**	240	-

Equal payment

Number of colours named	**Total income when you name red:**				**Total income when you name blue:**			
	red player named red	red player named blue	blue player named blue	blue player named red	red player named red	**red player named blue**	blue player named blue	blue player named red
3 red	**310**	-	-	240	310	**310**	-	240
2 red and 1 blue	**310**	310	240	240	310	**310**	240	240
1 red and 2 blue	**310**	310	240	240	310	**310**	240	240
3 blue	**310**	310	240	-	-	**310**	240	-

Zero payment

Number of colours named	**Total income when you name red:**				**Total income when you name blue:**			
	red player named red	red player named blue	blue player named blue	blue player named red	red player named red	**red player named blue**	blue player named blue	blue player named red
3 red	**250**	-	-	100	250	**250**	-	100
2 red and 1 blue	**250**	250	100	100	250	**250**	100	100
1 red and 2 blue	**250**	250	100	100	250	**250**	100	100
3 blue	**250**	250	100	-	-	**250**	100	-

D. Utility of the Green Player

Utility function (2) in Section 3 can be directly derived from the sum of Fehr-Schmidt utilities for the four active participants (two red and two blue players):

$$u_g = \sum_{i=1}^{2}(\pi_r^i + \pi_b^i) - \frac{\gamma}{3}\sum_{j=1}^{2}\sum_{i=1}^{2}\left|\pi_r^i - \pi_b^j\right| - \frac{\delta}{3}|\pi_r^1 - \pi_r^2| \tag{2}$$

A classical Fehr-Schmidt utility function has two parts, the own payoff of player *i* as well as disutility from inequality between i and the other players of the group:

$$U_i(\pi) = \pi_i - \alpha \frac{1}{N-1} \sum_{j \neq i} max\{\pi_j - \pi_i, 0\} - \beta \frac{1}{N-1} \sum_{j \neq i} max\{\pi_i - \pi_j, 0\}.$$

We denote the payoff of each red player of the group of four active participants with π_r^i , where $i = 1{,}2$ and the payoff of each blue player with π_b^i , where equally $i = 1{,}2$.[28]

When we sum up the Fehr-Schmidt utilities of the four active participants, the first part becomes just the sum of the payoffs of the players:

$$\pi_r^1 + \pi_b^2 + \pi_r^3 + \pi_b^4.$$

For the second part we need to consider all six possible differences between players:

$$\left|\pi_r^1 - \pi_b^1\right|, \left|\pi_r^1 - \pi_b^2\right|, \left|\pi_r^2 - \pi_b^1\right|, \left|\pi_r^2 - \pi_b^{j2}\right|, |\pi_r^1 - \pi_r^2|, \left|\pi_b^1 - \pi_b^2\right|.$$

When one of these inequalities is equal to zero, it can be neglected. This is always the case with the payoff difference of two blue players $\left|\pi_b^1 - \pi_b^2\right|$. When the difference is non-zero, it is always to the advantage of one player, but to the disadvantage of the other. Hence it needs to be considered twice. As an example look at $\left|\pi_r^1 - \pi_b^1\right|$ and assume that $\pi_r^1 = 370$ and $\pi_b^1 = 330$.[29] Then $\left|\pi_r^1 - \pi_b^1\right| = |370 - 330| = 40$. This inequality is to the disadvantage of the blue player and will therefore cause a disutility of $-(\frac{\alpha}{3}40)$ for him. For the red player, the inequality is to his advantage and causes therefore $-(\frac{\beta}{3}40)$. For the green player, the situation is symmetric, since all payoff differences are either zero or weighted with $\alpha + \beta$:

$$-\frac{\alpha+\beta}{3}(\left|\pi_r^1 - \pi_b^1\right| + \left|\pi_r^1 - \pi_b^2\right|, +\left|\pi_r^2 - \pi_b^1\right| + \left|\pi_r^2 - \pi_b^{j2}\right| + |\pi_r^1 - \pi_r^2|).$$

We now define $\alpha + \beta = \gamma$. In a last step we make a small but straightforward addition to Fehr-Schmidt. We single out $|\pi_r^1 - \pi_r^2|$, because the authority may perceive this inequality as more problematic and therefore add an extra weight $\varepsilon \geq 0$ to it, such that $\delta = \gamma + \varepsilon$:

$$-\frac{\gamma}{3}\left(\left|\pi_r^1 - \pi_b^1\right| + \left|\pi_r^1 - \pi_b^2\right|, +\left|\pi_r^2 - \pi_b^1\right| + \left|\pi_r^2 - \pi_b^{j2}\right|\right) - \frac{\delta}{3}|\pi_r^1 - \pi_r^2|).$$

[28] The actual numeration of the two red and two blue players is arbitrary.

[29] This is the case when the authority chose *max* and the “first” red player lied about his color while the “second” red player told the truth.