# Individual differences in social dilemmas: the effect of trust on costly punishment in a public goods game

*Esmee Bosma (5994411)*

*Supervisor: Prof. dr. ir. V. Buskens*

*Sociology, Utrecht University, The Netherlands*

*June 13, 2019*

**Abstract**

The establishment of cooperation in social dilemmas is important to real life problem solving and improving the environment, for example in a neighbourhood. Cooperation is promoted when subjects believe the interdependent others are able to punish. Individual characteristics of persons can be of influence on cooperation and punishment behaviour. This study focuses on individual differences in trust and investigates the effect of trust on cooperation and punishing behaviour in a linear public good game with peer punishment opportunities. The research question is: 'What is the effect of individual differences in trust on the likelihood of punishing non-cooperative behaviour of fellow players in public good games with punishing possibilities?'. Experimental data of 148 participants is used to research their punishment behaviour. Following the social reciprocity theory, expected is that more trust leads to more negative feelings if others have non-cooperative behaviour of, and therefore to more punishment. Multilevel regression is used to analyse the data. The results demonstrate a positive effect of trust on cooperation, yet the data contains no effect of trust on punishment. This suggests that punishment is possibly motivated by the contribution of other players rather than by trust, or that trust leads for some persons to more punishment and for others to less punishment. The role of trust in punishing behaviour remains uncertain, and future research can react to this by focussing on individual motivations and emotions during choice making in a social dilemma with punishment opportunities.

**Key words:** Social dilemma; costly punishment; trust; cooperation; public goods game; social reciprocity theory

**Word count:** 9390

# Table of contents

## 1. Introduction

In many of our daily life choices, there is a conflict between our personal interest and collective interest. Should you participate in activities that will help to protect the environment? For instance, should you have a shower for a shorter time, or should you separate the garbage in different dustbins? In these situations, the selfish choice is to follow your own personal interest (enjoying a longer shower) and not to contribute to the collective interest (saving the planet). The benefits for an individual are higher when he or she does not cooperate, than when the individual does. However, when everyone else also follows his or her selfish decision, the general outcome for everyone will be worse than in the case of the existence of some cooperation. These types of interdependent situations are referred to as social dilemmas (Dawes, 1980; Kopelman, Weber, & Messick, 2002; Liebrand, Messick, & Wilke, 1992).

Social dilemmas can take the form of public goods games. In public goods dilemmas, individuals decide how much they want to contribute for the maintenance or the improvement of a group fund or public good. This public good is a resource from which all individuals can benefit. This can be the refurbishment of a neighbourhood or the contribution to a group project (Gächter & Herrmann, 2009). All individuals can decide whether they contribute to this public good. However, no one is excludable from this resource: individuals who contributed to the public good are unable to prevent the non-contributors from consuming the good (Conybeare, 1984). As a result, the advantages of the public good are equally distributed over everyone, and all actors benefit whether they cooperated or not. People who don't contribute or contribute less than average, are called free riders (defectors). The lack of incentive for actors to cooperate is known as the free-rider problem. So, when only taking into account individual costs and benefits, the rational decision for an individual is to free ride. However, if everyone follows this decision, the public good is not provided.

Certain conditions can have a positive effect on the establishment of cooperation in a social dilemma. First, cooperation is promoted when actors believe the interdependent others are able to punish non-cooperative behaviour and/or to reward cooperative behaviour (Komorita & Barth, 1985; Komorita, 1987). These sanctions can be given by informal actors, for example by neighbours who point out to you that your garbage is not separated well. Also, cooperation can be increased when formal actors, like governments, provide social sanctions about an individual's behaviour. Those institutions can implement positive and negative sanctions to make cooperation more attractive or even force it. Second, an individual is more

likely to cooperate when the individual interacts in a social network whose members provide a form of social control (Raub & Weesie, 1990). In that case, the individual adheres to the norm: a behavioural rule which prescribes what is proper or improper behaviour in a certain context, and is aimed to promote positive action, e.g., cooperating (Hechter & Opp, 2001). Third, differences in group size influence cooperation. Cooperation decreases when groups become larger (Hamburger, Guyer & Fox, 1975; Bonacich, Shure, Kahan & Meeker, 1976), most of the time due to reduced feelings of responsibility (Olson, 2009) and reduced feelings of contribution to the collective goal (Fleishman,1980). When groups become larger than approximately seven/eight persons, the level of cooperation is not strongly influenced by differences in group size anymore (Fox and Guyer, 1977; Liebrand, 1984).

When an individual does not adhere to the norm, other individuals have the opportunity to sanction these defectors by punishing them. Negative emotions towards defectors compose an important mechanism behind punishment (Fehr & Gächter, 2002). Whether someone should punish or not, is a social dilemma in itself: a second-order free-rider problem (Heckathorn, 1989; Oliver, 1980; Yamagishi 1986). Punishing defectors is costly for individuals: actors have to spend resources to punish others. For example, it is possible to confront a group member with his/her low contribution to the group project. However, confronting this person costs time and energy. Therefore, everyone prefers others to invest in the punishment and does benefit from the results of punishment (probably less norm violation) at the same time. However, if everyone follows this strategy, punishment will not occur, and defection is still going to take place in the primary social dilemma. The result of punishment is that the level of cooperation rises (De Quervain, Fischbacher, Treyer, & Schellhammer, 2004; Fehr & Gächter, 2000; Gächter, Renner & Sefton, 2008; Gürerk, Irlenbusch & Rockenbach, 2006; Kroll, Cherry & Shogren, 2007; Rockenbach, Milinski, 2006).

Individual characteristics can play a role in decision making in a first-order (primary) dilemma, but also in a second-order (secondary) dilemma. Research about individual differences and first-order cooperative behaviour has already been conducted. This started with individual differences in social value orientation (SVO). Later, the attention was also focused on individual differences in trust (Liebrand, Messick & Wilke, 1992). Individual differences in these concepts influence the perceptions of social dilemmas and how people approach others. Trust is one of the most important factors for promoting and maintaining cooperation (Van Lange, Rockenbach & Yamagishi, 2017) and can be defined as: `The willingness of a party to be vulnerable to the actions of the other party, based on the

expectations that the other will perform a particular action important to the trustor (first person who places trust), irrespective of the ability to monitor or control that other party' (De Cremer, Snyder & Dewitte, 2001). Known is that people high in trust are more likely to cooperate than people low in trust (Liebrand, Messick & Wilke, 1992). Less research has been done about the relationship between trust and the second-order dilemma, while trust appears to be an important factor across many social interaction situations and is often described as 'social glue' to relationships, groups and societies (Van Lange, Rockenbach & Yamagishi, 2017). Therefore, this research will be about the relationship between trust and punishing behaviour. The research question is 'What is the effect of individual differences in trust on the likelihood of punishing non-cooperative behaviour of fellow players in public good games with punishing possibilities?'. Yamagishi (1986) found differences between groups of low and high trust and their contribution to the establishment of a sanctioning system. This research will be complementary, because it investigates the non-repeated interactions in the social dilemma between individuals with different levels of trust.

Understanding the willingness and knowing factors that contribute to engaging in costly punishment is a key element in understanding our sociality. Individuals do sometimes cooperate with the authorities in reporting illegal activity, and eyewitnesses accept to testify in favor of, or against, an unfamiliar person and don't expect any future benefits from doing so (altruistic punishment). The availability of costly sanctions has been shown to help enforce a social norm of cooperation among unrelated individuals (Fehr and Gächter, 2000). So, understanding the effect of trust on punishment leads to understanding the way in which social dilemmas can be solved. Logically, in our society, non-cooperative behaviour is not desirable. Parties can greatly benefit by cooperating in local projects and international politics, and goals can be reached which are impossible to reach individually. For example, the renovation of a community, or an international agreement for taking steps to protect the environment. Therefore, it can be profitable to obtain more knowledge about establishing first- and second-order cooperation, and about the individual differences in trust that play a role in choosing during the social dilemma.

## 2. Theory

In this section, past research regarding trust and punishment is described. First, general theories about cooperation and sanctioning are explained. These theories preceded to a new theoretical approach (in that time) of Yamagishi which was tested by means of experiments in 1986 and 1988. As far as I know, there is no other published research in which the effect of trust on punishing behaviour is tested. Second, a short outline of the development of theories concerning individual differences which affect cooperation in the first- and second-order social dilemmas is given. Third, social reciprocity theory is described, which states that sanctioning is driven by negative emotions towards norm violators. Last, the relation between negative emotions and the effect of trust on punishment is described and leads to a final hypothesis concerning individual differences in trust in combination with different group compositions in the social dilemma.

### 2.1 General theories from past research

An elementary theory applied to a social dilemma is the rational-structural approach (Olson, 1965). This theory uses the reasoning that people will choose in a self-interested manner: people will only contribute to a social dilemma when the personal benefit exceeds the personal cost. Therefore, the only solution to solve the social dilemma is a modification of the incentive structure, which causes a change in the individual pay-offs. An adjustment of the incentives can be realized by introducing positive or negative sanctions. As a result, the advantages of defection decrease for each individual. However, a sanctioning system as a structural change is a second-order public good. Following the rational-structural approach, the rational choice is not to contribute to the sanctioning system (the public good), because of the personal costs. According to this approach, a sanction system is not established, and cooperation never takes place.

After the rational-structural approach, the 'goal-expectation' approach was developed (Pruit and Krimmel, 1977). This approach assumes people to have a broader perspective than only considering their own short-term costs and benefits. Following this theory, most of the time people know about the disastrous consequences of a self-interested choice and know that cooperation can be essential for their own long-term benefits. Therefore, individuals develop a goal of mutual cooperation. Results of an experiment (N = 192) of Yamagishi (1988a) showed that subjects cooperate more when the gain for cooperation is larger: as the social dilemma becomes more serious, people become more willing to cooperate. Once a goal of

collaboration is developed, the decision to cooperate still depends on the trust in partners' willingness to reciprocate the cooperation, according to Pruitt and Kimmel (1977). Therefore, this theory is less applicable in large groups: individuals will feel like they have a big chance to be exploited and will therefore not cooperate.

In 1986, Yamagishi tested a new theoretical approach: the 'structural goal/expectation theory'. The structural goal expectation approach is a combination of the rational-structural approach and the goal-expectation approach. This new approach assumes that people who have developed the goal of mutual cooperation do cooperate for the implementation of a sanction system (instrumental cooperation), rather than engaging in cooperative actions in the original social dilemma. However, the trust that other members will cooperate, may prevent people from realizing the necessity of the structural change. Following this theory, establishing a punishment system will only take place if people realize the impossibility of voluntarily based cooperation and the importance of structural changes. This theory was tested by means of a resource experiment in Japan with 48 same-sex four-person groups, divided into groups of high and low in trusting other people. The first results, in a setting without a sanction system, showed that the level of cooperation of participants from the high-trust group was higher than the level of cooperation of the participants from the low-trust group. The results of a setting in which it was possible to establish a sanction system, indicated that people low in trust were more likely to develop a sanction system and contributed more to it than people high in trust (Yamagishi, 1986). This experiment was conducted again in the United States (Yamagishi, 1988b). Besides the fact that the general level of trust was higher in the United States, the results were equal.

Following the 'structural goal/expectation theory', people with low trust levels will contribute more to the establishment of a punishment system, because they realize the need of it. It is reasonable that also in a game in which a punishment possibility already exists, a structural motive for sanctioning only comes into force when the game is played several times with the same persons (repeated). Punishment is costly and will probably result in more cooperative behaviour of the opponent in the future. However, only when someone plays again with the same person the effects of punishment will have personal benefits. Therefore, this theory is probably less applicable in games with different opponents each round: individuals have no chance to play against the specific person again and will therefore probably not punish with the rational intention to obtain better behaviour in the future.

## 2.2 Individual differences in cooperating behaviour

In 1968, Messick and McClintock noticed that differences in individual motivational orientations underlie choices in social dilemmas. Since then, individual differences in social value orientation (SVO) are discovered (Messick and McClintock, 1968). Out of multiple orientations, three have received the most empirical and theoretical attention: (a) cooperation: the orientation to maximize own and others' outcomes; (b) individualism: the orientation to maximize one's own outcomes; and (c) competition: the orientation to maximize the relative advantage over others (Liebrand, Messick & Wilke, 1992). Ertan, Page & Putterman (2009) theorized that always choosing self-interested, which predicts absence of cooperation and absence of punishment, no longer holds because of these multiple preference types. Their experiment (N = 160) showed that 'the cooperators' are likely to punish low contributors because they dislike free-riding. Also, some subjects are 'perverse punishers', they have the preference to punish high contributors too. In a meta-analysis of 82 studies measuring the relationship between social value orientation (SVO) and cooperation in social dilemmas, Balliet, Parks, and Joireman (2009) show that over 40 years the individual differences in SVO have turned out to be a small but significant predictor of cooperation.

The research on social dilemmas focused besides individual differences in SVO, on individual differences in trust (Liebrand, Messick & Wilke, 1992). Trust is considered to be important in the context of general cooperative behaviour. Trust influences expectations about others' motives (Liebrand, Messick & Wilke, 1992) and is important for establishing cooperation (Yamagishi, 1986). People are more likely to cooperate if they expect others to cooperate as well, than if they expect others not to cooperate. When people have high levels of trust, people will have confidence in someone else's goodwill and moral behaviour, and therefore engage in reciprocal cooperation (Granovetter 1992). According to this theory and the past results described above (Yamagishi, 1986), it is reasonable that one's trust leads to a cooperative decision in a non-repeated public good game. The following hypothesis can be derived:

Hypothesis 1: People with high levels of trust will be more cooperative than people with low levels of trust.

## 2.3 Expression of negative emotions

In experimental situations as well as in real-life situations, it turns out that cooperation and sanctioning also take place in case an individual is not necessarily going to meet again with the norm violator in the future. Exchange does take place in a trust problem experiment whereby a buyer has to place trust in the seller and vice versa, and a self-organized reputation system in online markets does exist, even without state intervention (Diekmann & Przepiorka, 2017). Although it does take a little effort, buyers rate the sellers after the transaction and others can benefit from this information. Buyers give positive feedback to a seller when satisfied with the quality of the goods or with the quickness of the transaction process. Consistently, many buyers punish a seller who delivers poor quality goods by giving the seller a negative rating. These online market case studies show that the reputation mechanism can stimulate cooperative market transactions, as long as it is in all actors' own interest to observe certain rules (Diekmann & Przepiorka, 2017).

This is consistent with 'the social reciprocity theory' of Carpenter and Matthews (2004). This theory states that punishment can be explained by emotions, rather than by pay-off results or structural improvement ideas. Social reciprocity can be defined as: 'The act of demonstrating one's disapproval, at some personal cost, for the violation of a widely-held norm, regardless of the material consequences' (Carpenter and Matthews, 2002). In their experiment, individuals had the opportunity to sanction people in distinct groups and about half of the players did so. At the end of the experiment, participants were asked by means of a survey why they punished outside their groups. 86 percent punished 'to get back at rule breakers' and only 14 percent punished to increase the contributions in the other group. Based on these results, they conclude that social reciprocity is a robust theory of punishment and is the enforcement mechanism behind social norms.

As the questionnaire results of Carpenter and Matthews (2002) already show that sanctions are driven by negative emotions towards norm violators, the tendency to respond to positive actions positively and to negative actions negatively is clearly present in more experiments. Fehr & Gächter (2002) found that a free rider triggered much anger among the participants if these subjects contributed a lot relative to the free rider. Also, punishment increased when the deviation of the free rider from the average investment of the other members increased. Two years later, Fehr & Fischbacher (2004) found that a large percentage of people are willing to enforce cooperation norms, even though they incur costs, do not benefit from their sanctions and have not been directly harmed by the norm violation. De

Quervain, Fischbacher, Treyer, & Schellhammer (2004) studied brain activity while punishment was carried out. Findings show that people derived satisfaction from punishing norm violations. Moreover, results showed that there was more activation in the brain part that causes satisfaction when the monetary punishment was higher. Social reciprocity is also visible in the ultimatum game experiment of Xiao & Houser (2005). In an ultimatum game, participants reject or accept an amount of money which is offered by the opponent. The participants will decide if they think the offer is sufficient or insufficient and rejecting or accepting the offer is a way for the responder to display a reaction and sanction the proposer. In addition, in this experiment, half of the participants had the option to send a written message to the opponent. Results showed that participants with this option used the message to express emotions (Xiao & Houser, 2005). Masclet and Villeval (2006) theorized that in a cooperation game, negative emotions (anger, dissatisfaction) are non-strategic motives for individuals to punish others. The results of their experiment showed that negative emotions are a primary motive for punishment: participants did punish when there were no strategic reputation gains from doing so. Also, the intensity of punishment increased when the level of inequality increased and when personal earnings decreased.

## 2.4 Trust and punishment mediated by negative emotions

The social reciprocity theory can be applied to understand the possible effect of trust on punishment. As stated in 2.3, the social reciprocity theory argues that negative emotions are non-strategic motives for individuals to punish others. This implies that experiencing more negative emotions will lead to more punishment. It is arguable that people high in trust will experience more negative emotions than people low in trust, once their trust is abused by a decision of not cooperating of other subjects. This because people with higher trust levels have higher expectations about others' cooperation (Liebrand, Messick & Wilke, 1992). When others do not cooperate, these expectations do not come true and it is likely that the person with high trust will be hurt. A damaged trust leads to general emotional displeasure, and specific emotional reactions. For example, anger and fear (Tomlinson & Mryer, 2009). Therefore, it is assumed that the more trust someone has, the more negative feelings will be experienced during the experimental game. Those negative emotions can lead to more punishment. (Carpenter and Matthews, 2002; Fehr & Gächter, 2002; Fehr & Fischbacher, 2004; Quervain, Fischbacher, Treyer, & Schellhammer, 2004; Xiao & Houser, 2005; Masclet and Villeval, 2006).

Altogether, Yamagishi (1986) found that the lower the trust level among persons in one group is, the more likely it is they will contribute to developing a sanction system. However, these results are from an experiment with a similar group composition during the public goods game. Contribution to a sanctioning system is therefore personally profitable for the player because it will affect the behaviour of its group opponents and increases first-order cooperation. However, this research uses the data of an experiment with different group compositions in a public good game. Therefore, the structural goal/expectation theory is not applicable to this research. Playing with different persons each round makes it less likely that people punish out of rational thoughts for improvement of cooperation in subsequent rounds. Another difference is that the experiment of this research already contained a possibility for punishing. For these two reasons, it is assumed that punishment in the experiment will be driven by social reciprocity, rather than by rational thoughts. It is expected that people high in trust will experience more negative emotions due to norm violations than people low in trust, and will consequently punish more:

Hypothesis 2: People with high levels of trust will spend more points on punishment than people with low levels of trust.

# 3. Methods

## 3.1 Data

To test the hypotheses, the dataset of Quite (2013) is used. This data is from a computerized experiment, conducted in the Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University in the Netherlands. The experiment was a non-repeated linear Public Good Game (PGG) with a contribution and a punishment stage. In total 148 subjects participated, divided over six different sessions between December 2012 and January 2013. The participants were recruited from a subject pool of the laboratory. Most of these participants (85.8%) were students at Utrecht University. 66 of all participants (44.6%) were male and 96 (64.8%) of the participants were Dutch.

At the beginning of the experiment, all participants received identical instructions about the game. Participants were randomly divided into groups of four and the composition of the groups is different each round. In total, the game contained 18 rounds. In every round, the participant needed to first make a choice about his/her contribution to the public good, and second about his/her possible punishment to other group members. At the start of the Public Good Game, participants saw their own, and their group member's parameters on the screen. These were the number of points the participant had to start with (endowment = $X_i$), the received profit from the group project (demand = $D_i$), and the punishment effectiveness ($M_i$). [1] In the contribution round, each subject got the choice to contribute some of his/her points ($C_i$) to the public good. The public good is the sum of the contributions of the four group members, and the total profit is equal to all contributions × 2. The total profit was distributed over all group members divided by their demand. So, the earnings for a participant in the contribution stage of one round ($CI_i$) were: "participant's endowment" – "participant's contribution" + "participant's share" × "Total profit group project":

$$CI_i = X_i - C_i + D_i\left(\sum_{k=1}^{n} C_k \cdot 2\right) \qquad \text{(Quite, 2013)}$$

[1] Originally, the experiment was conducted to measure the effect of inequality. Therefore, the game contained three parts, each consisting of six rounds. In every part, the endowment ($X_i$ = 10 or 15 or 20), demand ($D_i$ = 0.17 or 0.25 or 0.33) and punishment effectiveness ($M_i$ = 2 or 3 or 4) could change for every player and were showed on the screen.

Contributing to the public good has a positive effect on total group income, while it decreases the income of the contributing participant:

$$2 \cdot D_i < 1$$

This creates a tension between the individual and group interest. The dominant and selfish strategy is not contributing to the public good. As a result, the group outcome is likely to be Pareto-inefficient: every player earns less than what was possible if everyone contributed.

Hereafter, the punishment stage is played. First, everyone in the group will see the contributions of each group member to the project and their own earnings from the first stage. Now the participants can decide whether they want to punish one or more of their group members. However, punishment is costly. The participants receive 3 points ($P_i$) to use for reducing the earnings of group members and can divide these points over the group members. If they do not use these points they will be added to their own earnings. This creates the second-order social dilemma: contributing to the punishment can have an indirect positive effect on total group income, while it decreases the income of the participant. The points assigned to group members are multiplied by the participant's personal punishment effectiveness ($M_i$). In the punishment stage, the income of the participant ($PI_i$) is: 3 – points spent on punishment – total points of punishment received from all group members:

$$PI_i = 3 - \sum_{k \neq i} P_{ik} - \sum_{k \neq i} M_k P_{ki} \qquad \text{(Quite, 2013)}$$

The total income of the participant is the sum of the income from the contribution stage and the punishment stage:

$$I_i = CI_i + PI_i \qquad \text{(Quite, 2013)}$$

The total number of points earned during the experiment is adjusted to a corresponding amount in money. The money the participants earned in total was paid out in cash at the end of the experiment. The participants earned on average €13,00, ranging from €8,00 to €18,50.

## 3.2 Measures

### 3.2.1 Independent variable

The independent variable is general trust. In the article of Yamagishi & Yamagishi (1994) general trust is defined as: 'General trust is a belief in the benevolence of human nature in general and thus is not limited to particular objects'. In other words, the independent variable is defined as trustfulness in strangers and society in general. Trustfulness because it is about one's own choice to place trust or not, and trust towards strangers because it is anonymous who the other players are in the experiment. Multiple dimensions of trust are measured by means of a survey with 15 different questions about trust at the end of the experiment. The items related to trust have been adopted from Yamagishi & Yamagishi (1994). For example, statements were 'The people I trust are those with whom I have had a long-lasting relationship' and 'In this society, one has to be alert or someone is likely to take advantage of you'. Participants could choose on an ordinal seven-point scale if they agreed or disagreed with the statements. The scale was presented as numbers ranging from 0 to 6, with the text 'totally disagree' and 'totally agree' on the endpoints respectively. Exploratory Factor Analysis is performed to reduce this data into a valid trust variable. In the statements, a distinction of knowledge-based trust (the sense of security in dealing with others with whom one has a long-lasting relation and whom one knows well) and general trust (trust in society/strangers) was noticed. This distinction origins from Yamagishi & Yamagishi (1994). Therefore, the factor analysis is performed with the expectation of 2 different factors of trust (general trust and relational trust) and the factors to extract is set to 2 in the analysis. Principal Axis factoring is used as the extraction method because a normality test shows significant non-normally distributed data. Direct Oblimin is used as an oblique factor rotation, which allows the factors to correlate. The Bartlett's test of sphericity was significant ($\chi 2$ (105) = 11623.89, $p < 0.001$), which indicates that the trust variables are related. The Kaiser-Meyer-Olkin measure was sufficient (KMO = .736), so it is likely that a limited number of underlying factors explains the variance in the variables. Therefore, it is appropriate to use the factor analysis on this set of variables. Results of the factor analysis show indeed a distinction in statements about general trust in strangers and trust in people you know well. The obtained pattern matrix is displayed in table 1, only items with factor loadings of above .30 are shown. The independent variable will be constructed based on the first factor 'general trust' and consists of 7 items. This factor has an Eigenvalue of 3.753 and accounts for 25% of the variance in the data. To construct the general trust variable, the items with a negative factor

loading (2, 3 and 7) are first recoded into three new variables, because these statements are in the opposite direction relative to the other statements. The three variables were recoded so that totally disagree as answer on these statements received a score of 6 instead of 0 on trust, and a totally agree-answer received a score of 0 instead of 6. Hereafter, a reliability analysis of the scale of the 7 trust items shows a sufficient Cronbach's Alpha (N = 2960, $\alpha = .770$, average inter-item correlation = .295). It also shows that the Cronbach's alpha is not going to increase if an item would be deleted. Finally, the trust variable is constructed by taking the mean of the scores on the 7 trust items, so the scale raises from 0 to 6 (descriptive statistics are displayed in table 2).

**Table 1. Pattern Matrix with Factor Loadings for Trust Items**

| Scale items | Factor | |
|---|---|---|
| | 1 | 2 |
| 1. Most people are basically honest. | .562 | |
| 2. No matter what they say, most people inwardly dislike putting themselves out to help others. | -.483 | |
| 3. People are always interested only in their own welfare. | -.494 | |
| 4. Most people are trustworthy. | .706 | |
| 5. In this society one does not need to be constantly afraid of being cheated. | .521 | |
| 6. I trust most people. | .696 | |
| 7. In this society, one has to be alert or someone is likely to take advantage of you. | -.571 | |
| 8. I trust a person I know well more than one whom I don't know. | | .519 |
| 9. Generally, a person with whom you have had a longer relationship is likely to help you when you need it. | | .784 |
| 10. The people I trust are those with whom I have had a long-lasting relationship. | | .626 |
| 11. I am trustworthy. | | .339 |
| 12. If I were going to buy a used car, I would feel more comfortable buying it from a salesperson whom a friend had introduced me to in person rather than from a salesperson who is a total stranger. | | .428 |
| Percentage of Variance | 25.02 | 13.37 |
| Eigenvalue | 3.753 | 2.006 |
| Cronbach's Alpha | .775 | .690 |

### 3.2.2 Dependent variables

The first dependent variable in the analyses is cooperation. Every round the participant needed to make a choice about how many points to contribute from his/her endowment to the public good. Cooperation is operationalized as the variable 'contribution': the points a subject contributed to the public good. This is a ratio variable. 0 points contribution is the smallest contribution possible, and 20 points contribution is the biggest contribution possible (table 2).

The second dependent variable is punishment. After the contribution decisions, the participants chose how many points they want to spend from their punishing points to the sanctioning of the three other players in the group. 14 of the 148 participants did not spend any point on punishment during the entire experiment. 2 participants spent 54 points during the complete experiment which is also the maximum punishment possible (18 rounds times 3 punishing points). Punishment is operationalized as the total punishment points the participant spent in one round, because then it is possible to see multiple decisions of a player, including the related contributions per decision. The variable 'points spent on punishment' is used, which is a ratio variable too. The smallest punishment contribution possible in one round was 0 points, and the maximum total punishment possible was 3 points.

### 3.2.2 Control variables

The data origins from an experiment which was conducted with the purpose of measuring the effect of inequality. To analyze the effect of trust on punishment, it is important to exclude this inequality component. Therefore, different control variables concerning inequality are added to the analyses. The first control variable is 'the endowment' a player started with. This variable has the values 10, 15 or 20. The demand from the public good is aligned with the endowment and therefore not added as a control variable too. The second control variable regarding inequality is 'the punishment effectiveness' and took the values of 2, 3 or 4. The variables 'endowment' and 'punishment effectiveness' are included as continuous variables in the analyses.[2] The third control variable is the binary variable 'unequal on size', in which 1 means that the subjects are in a group where players are unequal on size: some players will have a higher endowment and demand than other players in the group. The fourth control variable 'unequal on punishment effectiveness' is binary as well and indicates whether the

[2] When the categorical variables 'endowment' and 'punishment effectiveness' were included as dummy variables in the analyses, the coefficients showed a linear effect. Therefore, they are treated as continuous variables.

subject is in a group where players are unequal on punishment effectiveness or not. The last control variable with respect to inequality is 'opposite size and punishment' and shows if the subjects' own endowment and punishment effect are aligned or not. The dummy variable will have a value of 1 when the endowment is highest and the punishment effect is lowest for a player, and vice versa.

For the second analysis (testing hypothesis 2 about punishment) two extra control variables are included. First, the cooperation of the other players in the group is controlled. It is very likely that the punishing points assigned in one round depend on how much the group contributed. The variable contribution of the other players is constructed by subtracting the own contribution from the total group contribution in one round. The variable raises from 0 to 50. Second, 'contribution' is added as a control variable in this analysis and is about someone's own contribution. This can be of influence on the punishment decision, because it is possible that on average people will punish others less when they did not contribute themselves. This variable is the same as the cooperation/contribution (dependent) variable, and thus raises from 0 to 20.

**Table 2. Descriptive Statistics of Included Variables**

| | *M* | *SD* | Minimum | Maximum | *N* |
|---|---|---|---|---|---|
| *Independent variable* | | | | | |
| General trust | 2.95 | 0.874 | 0.14 | 5.14 | 2664 |
| *Dependent variables* | | | | | |
| Contribution | 9.96 | 5.355 | 0 | 20 | 2664 |
| Punishment | 1.09 | 1.257 | 0 | 3 | 2664 |
| *Control variables* | | | | | |
| Endowment | 15.00 | 4.083 | 10 | 20 | 2664 |
| Punishment effectiveness | 3.00 | 0.817 | 2 | 4 | 2664 |
| Unequal on size | .67 | | 0 | 1 | 2664 |
| Unequal on punishment effectiveness | .67 | | 0 | 1 | 2664 |
| Opposite size and punishment | .22 | | 0 | 1 | 2664 |
| Contribution of the other players | 29.89 | 8.803 | 0 | 50 | 2664 |

## 3.3 Analysis

All analyses were carried out using SPSS, version 25. To test the hypotheses, linear multilevel regression is used. As the participants played three parts of six rounds, one person makes 18 times a choice. Because of this, the cases in the data are repeated measures of one person: the observations are not independent. The individual's actions over time might be correlated with each other, because within one person there are certain characteristics which can explain their punishing behaviour. Actually, the person can be seen as a moderator, because the effect of trust on punishment depends on what personality one has. Data from individuals that are measured repeated have a two-level structure: measurements are nested within individuals. The measures are the first level, and the persons the second level. 'Regular' data analysis methods cannot be used, since those assume independence of observations. Therefore, multilevel regression analysis is used which controls for these clustered observations at higher level. The analysis contains a random intercept, to model that with a mean level of trust the expected mean punishment may vary between persons. The regression model can be written as:

$$y_{ij} = \beta_0 + \beta_1 X_{1j} + \ldots + u_{0j} + e_{ij}$$

Where $y_{ij}$ is the cooperation/punishment of one choice (*i)* of one person (*j)*, $\beta_0$ is the overall mean of cooperation controlling for other independent variables, $\beta_1 X_{1j}$ is the level of trust of one person, $u_{0j}$ represents the person-specific deviation from the intercept *j* on cooperation/punishment, and $e_{ij}$ is a residual error of one specific choice of one person.

# 4. Results

## 4.1 Repeated measures

The results of the multilevel analyses show a significant variation at the person-level. The degree of similarity can be indicated by the intraclass correlation. The intraclass correlation is calculated using the formula: $\rho$ = intercept variance / (intercept variance + residual variance). In the first analysis, the between-person (level 2) variance in cooperation is 17.505, and the within-person (level 1) variance is 11.288. Here, the intraclass correlation $\rho = 17.505 / (17.505 + 11.288) \approx 0.6079$, which indicates that approximately 60,79% of the total variation in contribution can be accounted for by the person who plays the game in the experiment. In the second analysis the between-person variance in cooperation is 0.673, and the within-person variance is 0.912, so the intraclass correlation $\rho = .673 / (.673 + .912) \approx 0.4246$. Nearly 42,46% of the variance in punishment can be attributed to differences between persons.

## 4.2 Hypotheses

The results from the multilevel analyses are displayed in tables 3 (results hypothesis 1) and 4 (results hypothesis 2). Expected was that people with high levels of trust will be more cooperative than people with low levels of trust. The first results within table 3 provide evidence that general trust is a significant predictor of cooperation ($p = .035$) and that this effect is positive. A one-point increase on the general trust scale (0-6) leads to an expected increase of 0.6 points on contribution. Since the contribution scale raises from 0 to 20, the effect is neither extremely large nor small: someone with the lowest trust score contributes around 4 points less than someone with the highest trust score.

Besides, the results show that most of the control variables have a significant effect on cooperation too. First, endowment has a positive significant effect on contribution: a one-point increase in endowment results in a .809 increase in contribution. This effect is relatively large, because endowment raises from 10 to 20. A participant with the highest endowment will approximately contribute 8 points more, compared to a participant with the lowest endowment. Second, punishment effectiveness has a small positive significant effect on contribution too: if the punishment effectiveness increases with one, the participant's contribution will increase with .224. Third, the control variable 'unequal on size' has a significant small positive effect. If a participant is in a group where players have unequal

endowments and demands, the individual might contribute about a half point more than if the players have equal endowments and demands. Fourth, if the subject's own endowment and punishment effect are not aligned, this will have a small significant negative effect on cooperation: the subject will contribute about 0.9 points less. Last, inequality in punishment effect does not make a significant change in contribution.

**Table 3. Coefficients from two Multilevel Linear Regressions of Contribution on selected independent variables.**

| | Model 1 – Random intercept only | | Model 2 – Random intercept with predictors | |
|---|---|---|---|---|
| Variable | B | SE | B | SE |
| Constant | 9.964*** | .350 | -4.705*** | 1.012 |
| General trust | | | .600* | .282 |
| Endowment | | | .809*** | .026 |
| Punishment effectiveness | | | .224** | .084 |
| Unequal on size | | | .495*** | .138 |
| Unequal on punishment effectiveness | | | -.085 | .144 |
| Opposite size and punishment | | | -.856*** | .174 |
| Number of parameters | 3 | | 9 | |
| *N* level 1 | 2664 | | 2664 | |
| *N* level 2 | 148 | | 148 | |
| Residual variance (level 1) | 11.288*** | .318 | 8.149*** | .230 |
| Intercept variance (level 2) | 17.505*** | 2.115 | 8.521*** | 1.045 |
| Intraclass Correlation Coefficient | .608 | | .511 | |

*Note: *p < .05 **p < .01 ***p < .001 (two-tailed tests)*

Second, expected was that people with high levels of trust will spend more points on punishment than people with low levels of trust. The level of general trust is not a significant predictor of punishment ($p = .626$). This means that this data shows that the points spend on punishment are not significantly affected by someone's level of general trust, so the null hypothesis cannot be rejected. Other variables do have a significant effect on punishment. First, a 1-point increase in endowment results in a significant .110 decrease in punishment. This effect is relatively very large since a participant with the highest endowment will approximately punish 1,1 points less than a participant with the lowest endowment. Second, if someone's endowment and punishment effectiveness are not aligned, this results in a significant .194 increase in punishment, which is a small to medium effect because punishment raises from 0 to 3. Third, the higher the contributions of the other players, the less points on punishment a player will spend: if the contribution of the other players increases with 1 point, punishment will significantly decrease with .038 points. So, in the case a group contributes 50 instead of 0 points, the amount of assigned punishment points will be 1,9 points smaller. Last, the effect of one's own contribution is significant and positive as expected: the more you contribute to the public good in one round yourself, the more you will punish in that round. An increase of 1 in someone's own contribution results in a .074 increase in punishment. This means that if someone contributes the maximum points possible (20), the punishment might be increased with around 1,5 points, relative to zero points contribution. Remarkable is that punishment effectiveness is not a significant predictor of punishment ($p = .069$), although it does approaches significance.

**Table 4. Coefficients from two Multilevel Linear Regressions of Punishment on selected independent variables.**

| | Model 1 - Random intercept only | | Model 2 - Random intercept with predictors | |
|---|---|---|---|---|
| Variable | B | SE | B | SE |
| Constant | 1.090*** | .070 | 2.867*** | .311 |
| General trust | | | .039 | .079 |
| Endowment | | | -.110*** | .010 |
| Punishment effectiveness | | | .047 | .026 |
| Unequal on size | | | -.068 | .043 |
| Unequal on punishment effectiveness | | | .009 | .045 |
| Opposite size and punishment | | | .194*** | .055 |
| Contribution of the other players | | | -.038*** | .003 |
| Own contribution | | | .074*** | .006 |
| Number of parameters | 3 | | 11 | |
| *N* level 1 | 2664 | | 2664 | |
| *N* level 2 | 148 | | 148 | |
| Residual variance (level 1) | .912*** | .026 | .792*** | .022 |
| Intercept variance (level 2) | .673*** | .084 | .667*** | .083 |
| Intraclass Correlation Coefficient | .425 | | .457 | |

*Note: *p < .05 **p < .01 ***p < .001 (two-tailed tests)*

## 5. Conclusion and discussion

The main goal of this research was to determine the effect of trust on punishment. The research question was 'What is the effect of individual differences in trust on the likelihood of punishing non-cooperative behaviour of fellow players in public good games with punishing possibilities?'. Two hypotheses regarding the effect of trust on cooperation and the effect of trust on punishment were tested by means of multilevel regression analysis on experimental data of 148 participants. Prior studies have noted the importance of trust to establish cooperation. The relevance of trust is again clearly supported by the current findings: the results show a significant positive effect of someone's general level of trust on someone's cooperative behaviour, which means that the first hypothesis is confirmed. Theories and past research seem to direct to an effect of trust on punishing behaviour too. However, the results of the analysis of the data used in this research lack a significant effect of trust on punishment, so the second hypothesis is rejected.

Since the empirical findings do not provide support for an effect of trust on punishing behaviour, different conclusions can be made. First, it could be the case that an effect does exist but is not found in this study, possibly due to limitations of this research. Future researches might still find an effect of trust on punishment. Second, a non-significant effect could mean that punishing behaviour is not related to the level of trust someone has at all. In this case, other factors than trust affect punishment. For example, in a quick decision, punishment can be determined almost completely by the cooperative behaviour of the other players. As the results showed, the contribution of other players has a significant negative effect on punishment. Additionally, it is possible that one's trust is not hurt by non-cooperative behaviour during the game, and negative emotions will stay out. Third, an insignificant effect of trust on punishment could mean that an effect does exist but is not visible, because the consequences of trust on someone's punishing behaviour can be negative for some persons and positive for others. Yamagishi (1986) found that a higher level of trust leads to less punishment, so maybe do some persons with a high trust score punish less because they believe the cooperation is going to be better in the next round. Since a negative effect of trust on punishment was also not visible, it is possible to theorize that at the same time other persons high in trust are hurt and do experience more negative emotions that lead to punishment.

The most important limitation of this study lies in the fact that the participant had a certain amount of punishing points to divide over all the players, instead of a certain amount

of punishing points to assign to each player separate. When dividing the points over the other players of the group in a round, the decisions are dependent. For example, if the subject punished one player of the group with all the available points, the other two players of the group received per definition zero punishment points. As a result, in this study, the total punishment per round was considered, instead of the punishment per player per round. It would have been an opportunity to look at the punishment per player per round because then it is possible to control for the contribution of this specific player. To overcome this issue in the future, the subjects should receive a certain amount of punishing points per player per round. Another possible limitation of this study concerns the measurement of trust. The survey with the questions regarding trust was conducted at the end of the experiment. Therefore, it is possible that the answers given were influenced by positive or negative experiences during the experimental game. If that is the case, the measured level of trust may differ from a person's overall level of trust.

To be more certain about the (absence of) effect of trust on punishment, it is important for future research to find out more about the individual motivations behind punishment. For example, different types of people could be disentangled by means of a neuroscientific study of social decision-making. Takagishi et al. (2009) researched the activation of the anterior insula and found that activation of this brain region is related to experiencing negative emotions as a result of inequity in a social dilemma. Conducting comparable research in combination with a measurement of trust of the participants makes it possible to know if negative emotions lead to more punishment and if (a part of) the participants with a high level of trust experience more negative emotions than the people with lower trust levels. Moreover, in the case that people high in trust indeed do differ in emotions and their punishing behaviour, future research could include questions about the motivation behind the punishment decision in the experiment. Answers to these questions can reveal if people high in trust who punish less, are driven out of the belief that cooperation is going to be better in the next round. In this way, it is possible to see if trust leads for some persons to a positive effect on punishment, and for others to a negative effect on punishment.

Based on this research, at least one can conclude that individual trust levels play a role in the decision about the contribution to a public good. So, a high level of trust in society among individuals can result in more cooperative decisions. This information can be relevant to for instance governments. Governments can possibly increase levels of general trust, for example by organizing volunteer events, where some people help others. Those events cause people to see that some people do want to help other people, and possibly cause the general

social cohesion within society to be increased. In turn, the increased general trust of individuals can lead to more contribution to public goods, for instance more renovations in a neighbourhood will be carried out. The effect of individual trust levels on punishing behaviour remains uncertain. It is highly possible that motivations for punishing, for example testifying against an unfamiliar person, are independent of trust. In that case, other factors are of influence on the decision to testify. For example, the choice can be determined by the perceived seriousness of the crime, or by the moral behaviour of the witness. Future research should be conducted to completely rule out a possible effect of trust on punishment.

## 6. References


Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations*, *12*(4), 533-547.

Bonacich, P., Shure, G. H., Kahan, J. P., & Meeker, R. J. (1976). Cooperation and group size in the N-person prisoner's dilemma. *Journal of Conflict Resolution*, 20, 687-705.

Carpenter, J. P., & Matthews, P. H. (2002) Social reciprocity. *Middlebury College Department of Economics Working Paper.*

Carpenter, J. P., & Matthews, P. H. (2004). Why punish? Social reciprocity and the enforcement of prosocial norms. *Journal of evolutionary economics*, 14(4), 407-429.

Conybeare, J. A. (1984). Public goods, prisoners' dilemmas and the international political economy. *International Studies Quarterly*, 28(1), 5-22.

Diekmann, A., & Przepiorka, W. (2017). Trust and reputation in markets. *The Oxford Handbook of Gossip and Reputation. Oxford: Oxford University Press, nn-nn.*

De Cremer, D., Snyder, M., & Dewitte, S. (2001). 'The less I trust, the less I contribute (or not)?' The effects of trust, accountability and self-monitoring in social dilemmas. *European Journal of Social Psychology*, 31(1), 93-107.

De Quervain, D. J., Fischbacher, U., Treyer, V., & Schellhammer, M. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53(5), 495-511.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980-994.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868), 137.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63-87.

Fleishman, J. A. (1980). Collective action as helping behavior: Effects of responsibility diffusion on contributions to a public good. *Journal of Personality and Social Psychology,* 38, 629-637.

Fox, J., & Guyer, M. (1977). Group size and others' strategy in a N-person game. *Journal of Conflict Resolution*, 21, 323-338.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, 322(5907), 1510-1510.

Granovetter, M. (1992). Problems of explanation in economic sociology. *Networks and organizations: Structure, form, and action*, 25-56.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The quarterly journal of economics*, 115(3), 811-846.

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*(5770), 108-111.

Hamburger, H., Guyer, M., & Fox, J. (1975). Group size and cooperation. *Journal of Conflict Resolution*, 19(3), 503-531.

Hechter, M., & Opp, K. D. (Eds.). (2001). *Social norms*. Russell Sage Foundation.

Heckathorn, D. D. (1989). Collective action and the second-order free-rider problem. *Rationality and society*, 1(1), 78-100.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362-1367.

Komorita, S. S. (1987). Cooperative choice in decomposed social dilemmas. *Personality and Social Psychology Bulletin*, 13(1), 53-63.

Komorita, S. S., & Barth, J. M. (1985). Components of reward in social dilemmas. *Journal of Personality and Social Psychology*, 48(2), 364.

Kopelman, S., Weber, J. M., & Messick, D. M. (2002). Factors influencing cooperation in commons dilemmas: A review of experimental psychological research. *The drama of the commons*, 113-156.

Kroll, S., Cherry, T. L., & Shogren, J. F. (2007). Voting, punishment, and public goods. *Economic Inquiry*, 45(3), 557-570.

Liebrand, W. B. (1984). The effect of social motives, communication and group size on behaviour in an N-person multi-stage mixed-motive game. *European Journal of Social Psychology*, *14*(3), 239-264.

Liebrand, W. B., Messick, D., & Wilke, H. (1992). *Social dilemmas: Theoretical issues and research findings*. Garland Science.

Masclet, D., & Villeval, M. C. (2006). Punishment, inequality and emotions.

Messick, D. M., & McClintock, C. G. (1968). Motivational basis for choice in experimental games. *Journal of Experimental Social Psychology*, 4, 1–25.

Oliver, P. (1980). Rewards and punishments as selective incentives for collective action: theoretical investigations. *American Journal of Sociology*, 85(6), 1356-1375.

Olson, M. (1965). The theory of collective action: public goods and the theory of groups. *Harvard University Press*, Cambridge.

Olson, M. (2009). *The logic of collective action* (Vol. 124). Harvard University Press.

Przepiorka, W., & Berger, J. (2016). The sanctioning dilemma: A quasi-experiment on social norm enforcement in the train. *European Sociological Review*, 32(3), 439-451.

Pruitt, D. G., & Kimmel, M. J. (1977). Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology*, 28(1), 363-392.

Quite, Wouter (2013). The effect of inequality on public-good provision: Norm conflicts & asymmetric enforcement of cooperation. *Working paper*.

Raub, W., & Weesie, J. (1990). Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology*, 96(3), 626-654.

Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, *444*(7120), 718.

Simpson, B., & Willer, R. (2015). Beyond altruism: Sociological foundations of cooperation and prosocial behavior. *Annual Review of Sociology*, *41*, 43-63.

Takagishi, H., Takahashi, T., Toyomura, A., Takashino, N., Koizumi, M., &Yamagishi, T. (2009). Neural correlates of the rejection of unfair offers in the impunity game. *Neuroendocrinology Letters*, 30(4), 496-500.

Tomlinson, E. C., & Mryer, R. C. (2009). The role of causal attribution dimensions in trust repair. *Academy of Management Review*, *34*(1), 85-104.

Van Lange, P. A., Rockenbach, B., & Yamagishi, T. (Eds.). (2017). *Trust in social dilemmas*. Oxford University Press.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, 102(20), 7398-7401.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51(1), 110.

Yamagishi, T. (1988a). Seriousness of social dilemmas and the provision of a sanctioning system. *Social Psychology Quarterly*.

Yamagishi, T. (1988b). The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly*.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129-166.