Heterogeneity in Distributional Preferences: An Experimental Study

Hyoji Kwon

Graduate School of Economics
Waseda University

Doctoral Thesis

November, 2023

Acknowledgements

I extend my heartfelt gratitude to Professor Yukihiko Funaki for his invaluable guidance throughout the completion of this doctoral thesis. In moments of uncertainty and confusion as a researcher, Professor Funaki consistently steered me in the right direction, offering a wealth of opportunities and experiences that ultimately facilitated the realization of this thesis. I consider myself extremely fortunate to have collaborated with Professor Funaki, a respected researcher actively expanding both his research scope and network. I aspire to emulate his commitment to engaging with emerging researchers and embracing innovative ideas. I want to express my deep thanks to Professor Funaki for his mentorship in shaping my identity as a researcher.

I would also like to express my sincere thanks to Professor Jung-Kyoo Choi for his uplifting encouragement in Korea. My deep involvement with the field of economics began with Professor Choi's captivating lectures and discussions during my undergraduate and master's years. His invitation to explore the world of game theory, experimental and behavioral economics, accompanied by stimulating questions and guidance on human behavior and society, laid the foundation for an enduring interest in economics. Professor Choi's encouragement to take on the challenge of studying abroad opened up a new dimension in my academic journey. His consistent support and insightful advice have been instrumental in the successful completion of this thesis.

Expressions of gratitude are extended to Professor Kazumi Shimizu, for his thorough support as my associate advisor, providing detailed and relevant advice for the completion of this thesis. I am grateful to Professor Kenju Kamei for his intellectually stimulating research and invaluable guidance in advising my doctoral thesis. His kindness and profound knowledge were pivotal in its accomplishment. Special acknowledgment goes to Professor Yukio Koriyama of Ecol Polytechnic and Professor Charles Noussair of the University of Arizona for thoroughly considering my research, offering warm and appropriate advice, along with enlightening details and pleasantly surprising insights.

I take this opportunity to express my appreciation to my colleagues in Funaki's lab – Taro Shinoda, Ayano Nakagawa, Xin Fang, Daeseok Kim, and Hideaki Minami. Their collaboration, assistance in conducting experiments, and collegial advice have been indispensable.

Last but certainly not least, I am deeply grateful to my beloved parents and brother. Their unwavering support during the long journey of study, even in challenging circumstances, has been the cornerstone of my journey. Their steadfast encouragement and understanding of my choices have empowered me to stay true to myself, purely focus on my research interests, and lead a sincere and fulfilling life. Their love and guidance have played a crucial role in my accomplishments.

Hyoji Kwon
November 2023
Osaka

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Distributional preferences, which involve the consideration of inequality and one's ranking of payoffs within a group, play a crucial role in social decision-making. Individuals often demonstrate a willingness to incur costs to reduce the gap in payoffs and promote a more equitable distribution of resources(Charness & Rabin, 2002; Engelmann & Strobel, 2004).

Notably, two prominent models— the inequality aversion model proposed by Fehr and Schmidt (1999) and the ERC (Efficiency, Reciprocity, and Competition) model introduced by Bolton and Ockenfels (2000)— incorporate distributional preferences into individual utility functions. These models explain behaviors such as contributions to public goods, the costly punishment of free-riders, and altruistic allocations in the ultimatum and dictator games. Experimental findings have demonstrated variations in distributional behavior in the ultimatum game and the dictator game across countries, cultures, family income levels, gender, and age. Based on this evidence, numerous studies have actively investigated the factors influencing people's distributional choices (Benenson, Pascoe, & Radmore, 2007; Brañas-Garza, 2006; Chen, Zhu, & Chen, 2013; Dreber, Ellingsen, Johannesson, & Rand, 2013; Heinz, Juranek, & Rau, 2012; Henrich et al., 2001; Krupka & Weber, 2013).

In Chapter 2, I emphasize the significance of inequality aversion as a driving factor for punishment in public goods games, as demonstrated by Fehr and Schmidt (1999). Their model suggests that inequality aversion leads to substantial levels of contribution and establishes a norm for appropriate contributions. According to their model, individuals are motivated to punish behavior that deviates from this norm. However, since punishment in public goods games serves a dual purpose of penalizing undesirable behavior and reducing inequality between cooperators and free-riders, explanations of motivations for punishment need to encompass both purposes. In other words, the explanations provided by Fehr and Schmidt (1999) only cover the first purpose, which is penalizing undesirable behavior.

Our first experimental study in Chapter 2 aims to distinguish between these two motivations and highlight the importance of inequality aversion as a motivation for punishment in public goods games. To differentiate between these motivations, I introduce a modification that incorporates uncertainty. By weakening the direct link between behavior and payoffs through the inclusion of an uncertainty factor, individuals can independently choose to punish behaviors that deviate from their cooperative norm and to reduce the gap in payoffs. Through this modified game, I identify three distinct types of individuals based on their motivations for punishment: reciprocal types, inequality-averse types, and an 'other' type without consistent punishment patterns.

Importantly, individuals exhibit heterogeneity in their motivations, with a distinct

group solely driven by inequality aversion. Additionally, even among individuals classified as the reciprocal type, punishment behavior based on inequality aversion is observed, although it is weaker than reciprocity. This characterization of different types in motivations for punishment highlights the significance of distributional preferences for both inequality-averse punishers and individuals prioritizing reciprocity.

Chapter 3 explores plural distributional preferences through dictator games involving real effort tasks. In situations where the endowment is jointly produced, and the magnitude of the endowment is determined by both individual effort and luck, determining who *deserves* what share and what constitutes a fair share become crucial questions. Additionally, if individuals lack distributional preferences and solely aim to maximize their own payoffs, they would choose the option that provides the highest monetary gain – taking all the shares.

Previous research by Cappelen, Hole, Sørensen, and Tungodden (2007) suggests that everyone possesses distributional preferences and adheres to different fairness ideals. Three distinct ideals emerge: strict egalitarianism, which aims for equal outcomes regardless of individual choice or luck; libertarianism, which emphasizes individual responsibility for income, including the results of luck; and liberal egalitarianism, which bases fair distribution on individual choices or efforts within reasonable influence. While acknowledging the existence of individuals who consistently maximize their own payoffs, our aim is to categorize fairness ideal types using our modified estimation model. Consequently, a notable proportion comprises pure egoists—individuals driven by self-interest and the pursuit of maximizing monetary payoffs. Even upon incorporating this distinct pure egoist type into our modified model, we emphasize the ongoing coexistence of these egoists with subjects adhering to the three fairness ideals proposed by Cappelen et al. (2007).

In Chapter 4, I further explore the characteristics of each fairness ideal type by conducting an additional experiment that incorporates an empathy treatment. Empathy, a crucial factor in fostering prosocial behavior, is examined in terms of state empathy, which is induced by specific situational factors. Our findings demonstrate that individuals with egoistic ideals exhibit a significant reduction in their share in response to state empathy, whereas individuals with fairness ideals show a less pronounced decrease in their distribution decisions in the empathy treatment. Essentially, state empathy primarily influences individuals who do not have specific fairness ideals, suggesting that its presumed effectiveness and widespread use in empathetic campaigns may be limited to a particular segment of the population. Furthermore, individuals with fairness ideals and distributional preferences prioritize maintaining their ideal distribution status over invoking empathy for the disadvantaged. These results provide valuable understanding of the complex interplay among empathy, fairness ideals, and distributional preferences, emphasizing the need to consider individual heterogeneity when designing interventions to promote fair outcomes.

In conclusion, this thesis explores heterogeneity in distributional preferences through an experimental study. By examining punishment motivations and fairness ideal types, and considering the impact of empathy, I uncover the complexities underlying decision-making processes related to distribution. The findings contribute to our understanding of how individuals perceive and act upon distributional issues, providing valuable insights for policymakers and researchers striving to foster fairer outcomes in society.

Finally, each chapter of this thesis is composed of the following papers:

Chapter 2. Kwon and Funaki (2023a), Heterogeneity in the motivations for punishment: An experimental study. *Available at SSRN 4441815.*

Chapter 3. Kwon and Funaki (2022), Do strict egalitarians really exist? *WINPEC Working Paper Series*, E2206 , 1–27.

Chapter 4. Kwon and Funaki (2023b), The empathetic egoist: The effect of empathy on plural fairness ideals. *Available at SSRN 4441637.*

1.2 Related Literature

1.2.1 Reciprocity and Inequality Aversion as Motivations for Punishment

Reciprocity has long been studied as the most important motivation for punishment (Fehr & Gächter, 2000, 2002; Andreoni, Harbaugh, & Vesterlund, 2003; Rockenbach & Milinski, 2006; Denant-Boemont, Masclet, & Noussair, 2007; Carpenter & Matthews, 2012). Reciprocity is a behavioral response to perceived kindness and unkindness, where kindness comprises both distributional fairness as well as fairness intentions. This motivation leads people to punish unkind and selfish free-riding behaviors. In other words, in reciprocal punishments, the intentionality of the undesired behavior is the most important consideration. Rabin (1993), Levine (1998), Dufwenberg and Kirchsteiger (2004), and Cox, Friedman, and Gjerstad (2007) suggest that fairness intentions and kindness matter in game models. In the context of experimental studies, reciprocal punishments have been explored by establishing a situation in which the ratio of a punishment's cost to its impact is 1:1. Since punishments cannot reduce the gap in payoffs, this setting can rule out the possibility that inequality aversion is the underlying motivation. Falk, Fehr, and Fischbacher (2005) conduct two kinds of three-player, one-shot prisoner dilemma games, one with the low sanction of a 1:1 cost–impact ratio and one with the high sanction of a 1:3 ratio, which is more effective than the 1:1 ratio. The authors show that the punishments inflicted by cooperators are not significantly different across conditions; thus, the reciprocity motivation is a better explanation for the observed punishments than inequality aversion. Furthermore, Egas and Riedl (2008) and Nikiforakis and Normann (2008) vary the effectiveness of the punishments and find that under the 1:1 punishment condition, reciprocal punishments always exist to a considerable degree in public goods games with costly punishments. However, the more effective the punishments are, the more punishments that are inflicted and the stronger those punishments are.

Masclet and Villeval (2008) compare public goods games with partner matching and with stranger matching under the 1:1 punishment condition. As the other studies discussed above have shown, a substantial number of punishments are imposed, and there is no significant difference in that number across matching protocols. These studies indicate that punishments are motivated by reciprocity rather than inequality aversion. In particular, however, I note the results of increased punishments under the 1:3 condition rather than the 1:1 condition. There are two possible reasons for this result. First, a relatively cheaper punishment could simply encourage more punishment in the 1:3 condition than in the 1:1 condition. This explanation does not weaken the argument in favor of reciprocal punishment. Second, only the 1:3 condition can reduce inequality, so this provides a clue regarding inequality aversion as a motivation, and I focus on this possibility. Since the more effective the punishments are, i.e., the more the punishments reduce payoff inequality, the higher the level of punishments induces, inequality aversion plays some role in the motivation to punish.

In addition to the prisoner's dilemma game and the public goods games, Falk, Fehr, and Fischbacher (2008) explore the motivations for punishment by conducting moonlighting games with an intention treatment and a no-intention treatment. This game consists of two stages with two players. In the first stage, Player A chooses an action from the choice set. If Player A chooses a negative number from the choice set, Player A takes as much of Player

B's endowment as the number he or she chooses indicates. If Player A chooses a positive number, Player A gives his or her endowment to Player B. In the second stage, Player B chooses to sanction or reward Player A's action. As a result, in the intention treatment, Player B punishes Player A when Play A takes Player B's endowment and gives a reward when Player A gives his or her endowment. However, in the no-intention treatment, in which Player A's decision is assigned randomly, Player B rarely chooses to sanction or reward. In this study, the authors reject the inequality aversion model (Fehr & Schmidt, 1999), which posits that people prefer an equal distribution regardless of intentionality. They suggest that fairness intentions (Falk & Fischbacher, 2006), similar to reciprocity, discussed above, matter for Player B's decision. This study also shows that people punish not the gap between payoffs but the intentions behind the behaviors that cause inequality. However, in a moonlighting game, retaliation may be harsher than that in a public goods game because taking others' endowments is an undesired and actually proactive behavior, similar to stealing. Unlike decision-making in a moonlighting game, decision-making in a public goods game involves the choice of how to distribute one's endowment to one's private pocket and to a public pot, so it is different from moonlighting games in which subjects proactively choose whether to take others' endowments. Moreover, selfish behaviors in a public goods game, namely, making no contribution, decrease the group's total payoff, but it is reasonable as an individual-level choice. In other words, in a public goods game, selfish behavior may not be intended to spite others but rather to maximize one's own profit, so punishment in a public goods game leaves room for another motivation, e.g., inequality aversion. Therefore, the situation in a moonlighting game creates the best situation for reciprocal punishments as revenge, but it cannot allow researchers to reject the possibility of inequality aversion in general punishment decisions.

On the other hand, inequality aversion focuses on payoffs rather than behaviors. In this model, which was suggested by Fehr and Schmidt (1999), people resist inequality and are willing to pay to reduce the gap between payoffs. Bolton and Ockenfels (2000) also propose that since people are motivated not only by monetary payoffs but also by relative payoffs, the equality of shares may maximize utility. To verify the existence of inequality aversion as a motivation for punishment, Dawes, Fowler, Johnson, McElreath, and Smirnov (2007) and Johnson, Dawes, Fowler, McElreath, and Smirnov (2009) design a random income game that skips the contribution stage of a public goods game and assigns payoffs following the empirical distribution of the payoffs in public goods games. After that, the punishment stage begins; subjects decide whether to punish the other players. Participants, of course, react to the inequality in payoffs, and the more others get, the more harshly they punish. This result shows that punishments can be used to reduce inequality. Zizzo (2003) reports that in money burning games, 75% of burners (equivalent to the punishers in other games) are rank egalitarians who adjust the level of the burning according to the rank of the others' payoffs. Since the money burning decision is based on the results of individuals' betting in the first stage, there are no interactions and there is no room for reciprocity. Hence, this study also points to inequality aversion as the motivation for punishment. Raihani and McAuliffe (2012) shows that when initial endowments are assigned unequally and disadvantaged players have an opportunity to cheat against advantaged players, the advantaged players rarely punish the disadvantaged players for cheating, even if they could do so, when the disadvantaged players cheat to equalize payoffs. However, when initial endowments are equal, approximately half of cheated players inflict costly punishments. From this result, the authors argue that people do not punish cheating behaviors reciprocally but punish the inequality caused by cheating. On the other hand, when inequality exists initially, cheating can be justified and is not

regarded as an unkind or selfish behavior. In other words, the choice to not punish may be motivated by reciprocity. Thus, responses to such cheating are not able to totally exclude reciprocity as a motivation, and it is difficult to conclude that reciprocity is not a motivation.

In our literature review thus far, reciprocity has been emphasized as a stronger motivation for punishment in various games than inequality aversion. However, in studies of inequality aversion as motivation for punishment, researchers have established situations that minimize or eliminate the effect of reciprocity so that inequality aversion is clearly what is at work and the researchers can concentrate specifically on the existence of inequality aversion. Moreover, it is possible that other social preferences affect punishment decisions. These preferences include, for example, efficiency maximization (Charness & Rabin, 2002; Engelmann & Strobel, 2004), envy (Casal, Güth, Jia, & Ploner, 2012; Zizzo & Oswald, 2001), competitiveness (Bault, Coricelli, & Rustichini, 2008), and spite (Hilbe & Traulsen, 2012).[1] Therefore, in Chapter 2, to distinguish reciprocity and inequality aversion as motivations for punishment, I design modified public goods games with costly punishment that introduce an uncertainty factor. Since this uncertainty factor allows the two motivations to be distinguished, our results suggest that people who hold each of the two motivations coexist within the population. Furthermore, I indicate that there are cases where people control the intensity of punishment by considering the two motivations together—that is, that the two motivations complement each other.

1.2.2 Experiments for Pluralism of Fairness Ideals

Fairness ideals are pluralistic. Cappelen et al. (2007) show the coexistence and pluralism of fairness ideals through dictator games preceded by a production phase (Konow, 2000). They assume that an individual favors one of three fairness ideals: strict egalitarianism, libertarianism, or liberal egalitarianism. Strict egalitarianism insists that all inequality should be eliminated and seeks equality of outcomes regardless of individual performance or productivity, as in the models of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). Libertarians, whose views are philosophically based on Nozick (1974), place individual ownership first, claiming that people are responsible for all factors affecting their income, even if some factors are shaped by luck. Thus, each individual has the right to own his or her outcome. The liberal egalitarian fairness ideal, which is similar to Konow (2000)'s accountability principles, asserts that fair distribution should be based on an individual's choice or efforts while excluding the role of those that he or she cannot reasonably influence, that is, randomly assigned factors (e.g., a physical handicap). Cappelen et al. (2007) establish a choice model that describes the tradeoff between the ideal offer in the dictator game according to each of three fairness ideals and selfishness. Then, they estimate the distribution of outcome corresponding to the three fairness ideals and show the pluralism of fairness ideals followed by their experimental participants.

Different fairness ideals coexist in society, but the proportions of those espousing these fairness ideals are not constant in the population. Various experimental studies examined which socioeconomic factors influence the distribution of adherence to the three fairness principles by means of the experimental design and choice model of Cappelen et al. (2007). Almås, Cappelen, Salvanes, Sørensen, and Tungodden (2017) observe an effect of family background on the subject's choice of a fair distribution. They show that the proportion of liberal egalitarians was significantly higher than that of strict egalitarians in a group

[1] We discuss in further detail the effect of social preferences other than reciprocity and inequality aversion in Section 2.4.

with high parental income and education levels. Cappelen, Moene, Sørensen, and Tungodden (2013) show that the predominant fairness principle differs between high-income and low-income countries. High-income countries have a much higher percentage of liberal egalitarians than of strict egalitarians, but in low-income countries, the percentages of strict egalitarians are higher than those of liberal egalitarians. Almås, Cappelen, Sørensen, and Tungodden (2010) conduct an experiment on children and adolescents, showing that the distribution of adherence to fairness ideals varies according to human development, that is, with age. As the grade increases, the proportion of strict egalitarians decreases, and the proportion of liberal egalitarians increases. These studies add a real-effort task in the production phase to Cappelen et al. (2007)'s experiments, so each subject's actual efforts determine the size of his or her group's endowment to be distributed. These studies suggest that the distribution of adherence to fairness ideals in the population may differ depending on the situation in each society. In general, it seems that the higher education and average income are, the more people choose the fairness ideal that values individual efforts or performance, namely, liberal egalitarianism.

Research on the factors affecting the distribution of adherence to fairness ideals in the population is important because it can help explain different choices of or conflicts over distribution policies across and within societies and organizations. However, studies based on Cappelen et al. (2007)'s estimation model are subject to critique. Essentially, the experiments in these studies are dictator games (Forsythe, Horowitz, Savin, & Sefton, 1994). In the dictator game, the receiver has no right to reject the dictator's offer, so the dictator's dominant strategy is to take all of the endowment. Although other previous experimental studies that include a production phase with real-effort tasks show that the dictator increases the receiver's share according to the latter's performance (Becker, 2013; Erkal, Gangadharan, & Nikiforakis, 2011; List, 2007; Mittone & Ploner, 2012; Ruffle, 1998), over 20 percent of dictators still make the selfish choice, taking everything. However, in any studies based on the experimental and estimation method of Cappelen et al. (2007) above, the proportion of strict egalitarians never falls below 20 percent, and it seems that no subjects make selfish choices. In Cappelen, Hole, Sørensen, and Tungodden (2011), this point is revealed indirectly. They compare the self-reported data on subjects' adherence to fairness ideals and the estimation results. In the cases of the libertarian and liberal egalitarian fairness ideals, the proportions of subjects who follow each ideal in the preexperimental questions and the estimated proportions of adherents to the two ideals found in the experimental data are not significantly different. However, in the case of strict egalitarianism, less than 10 percent of subjects answer that they prefer this ideal in the preexperimental questions, but more than 20 percent of subjects are estimated to be strict egalitarians from the experimental data. Rodriguez-Lara and Moreno-Garrido (2012) and Ubeda (2014) also criticize Cappelen et al. (2007)'s results, arguing that few subjects consistently follow the specific fairness principle, but, rather, people tend to follow different principles in self-serving ways. People choose the fairness principles that give them higher payoffs depending on their performance and randomly assigned productivity. Therefore, they conclude that fairness ideals seem to be context dependent.

In Chapter 3, against this background, I attempt to more accurately estimate the distribution of adherence to fairness ideals. Our study is based on the works of Cappelen et al. (2007) and Almås et al. (2010), with the addition of some modifications to the estimation models. In the first modified model, I introduce a pure egoist—that is, individuals who prefer to take the entire product for themselves—into the choice model proposed by Cappelen et al. (2007). In the second modified model, I additionally assume that the

parameter representing the weight on the fairness ideal (β in Cappelen et al. (2007)) can reveal the characteristics of each fairness ideal. Through these parameters, it is possible to compare how important each group of subjects following the same fairness ideal considers their ideal to be. I explain the details in Chapter 3.

1.2.3 State Empathy and Its Effect for Prosocial Behaviors

Empathy is a reaction to the observed experiences of another (Davis, 1983); it was first introduced in the 1990s (Rifkin, 2009) and studied extensively in the fields of psychology and neuroscience (Batson, 2009; Cuff, Brown, Taylor, & Howat, 2016; Kirman & Teschl, 2010; Singer & Lamm, 2009; Zaki & Ochsner, 2012). Adam Smith, quoted in the head of this chapter, refers to this emotion as sympathy both arising from the imagination of oneself in another person's shoes and generated by observing another person's feelings. However, according to a recent psychological definition, the former is closer to empathy and the latter is closer to sympathy (Batson, 2009; Cuff et al., 2016). Moreover, the term of empathy has been used for various concepts (Batson, 2009; Cuff et al., 2016; Preston & De Waal, 2002).

In particular, Cuff et al. (2016) review the definitions of empathy from 43 works of literature and similar to Batson (2009), explain that there are various aspects of empathy. In summary, they give the following definition of empathy: "Empathy is an emotional response (affective) dependent upon the interaction between trait capacities and state influences. Empathic processes are automatically elicited but are also shaped by top-down control processes. The resulting emotion is similar to one's perception (directly experienced or imagined) and understanding (cognitive empathy) of the stimulus emotion, with recognition that the source of the emotion is not one's own". In particular, Cuff et al. (2016) define *trait empathy* as an individual's "capacity", distinguishing it from *state empathy*, which is expressed in a specific "context" or "situation". According to Cuff et al. (2016), trait empathy implies that some individuals are more empathetic toward others, and this ability remains stable over time. On the other hand, state empathy is promoted by situations or conditions, such as the similarity between an observer and his or her target, the perception of the observer's power, and the cognitive load. Therefore, Cuff et al. (2016) conclude that empathy is a result of the interaction between state and trait influences.

Previous studies show that promoting affective and cognitive empathy, which are kinds of trait empathy as an individual ability, by presenting another's emotional state or situation significantly increases prosocial behavior, including altruistic giving and fairness considerations (Andreoni & Rao, 2011; Andreoni, Rao, & Trachtman, 2017; Basil, Ridgway, & Basil, 2008; Batson et al., 1991, 1997; Christian & Alm, 2014; Chuan & Samek, 2014; Edele, Dziobek, & Keller, 2013; Herne, Hietanen, Lappalainen, & Palosaari, 2022; Wang et al., 2022). However, relatively few studies focus on the effect of state empathy, which is promoted by providing a *situation*, namely, allowing a person to put himself or herself in another's shoes. Therefore, I examine the effect of promoting state empathy in a dictator game and show that the degree of the effect of state empathy differs based on subjects' fairness ideals by classifying subjects' fairness ideal types.

State empathy has been used to describe others' situations and to present sentences such as "Imagine that you are in this person's situation." Batson et al. (1991), Basil et al. (2008), Christian and Alm (2014), Herne et al. (2022), and Andreoni and Rao (2011) observe changes in prosocial behavior by promoting state empathy in this way. First, by conducting psychological experimental studies in which a subject's decision-making is not related to monetary incentives, Batson et al. (1991) and Basil et al. (2008) research

the effect of state empathy. They compare people's charitable donation intentions in low-empathy conditions and high-empathy conditions. In both studies, after explaining the situation of a person in need of help, in the high-empathy conditions, they present a sentence such as "Imagine that you are in this person's situation" to induce state empathy. The percentage of people who show empathetic responses to donations or charity activities is significantly higher in the high-empathy conditions than in the low-empathy conditions, in which only the situations of the people in need are presented and intentions to donate or engage in charity activities for a person in need are asked for. In addition, Basil et al. (2008) measure subjects' guilt and maladaptive responses through a survey in each condition, showing that state empathy works as a mediator to increase guilt and reduce maladaptive responses. Neither of these psychological experimental studies uses the term state empathy, but they promote empathy by presenting a specific situation in a high-empathy condition and inducing subjects to put themselves into another's situation. These high-empathy conditions can be considered to promote the same type of state empathy as that defined by Cuff et al. (2016).

Second, in the experimental economics approaches, Christian and Alm (2014), Herne et al. (2022), and Andreoni and Rao (2011) examine the influence of state empathy on prosocial behaviors. These experiments are monetarily incentivized experiments in which payoffs are determined by participants' decision-making. Christian and Alm (2014) distinguish empathy from sympathy, showing that promoting empathy has a positive effect on tax compliance. First, regarding empathy, they quote the definition of Batson, Coke, et al. (1981), that is, "an affective state of 'putting yourself in someone else's shoes', in which an individual feels the same or a similar emotion as the other person"; moreover, sympathy is defined as "an emotional response of sorrow or concern for another's wellbeing caused by the other's emotional state, a response that is not identical to the other's emotion". The phrase, 'putting yourself in someone else's shoes', is a representative expression referring to state empathy. In their experiment, they use the 'priming' method. This method consists of two parts; firstly, subjects read six versions of the Golden Rule[2] from different religions, and secondly, they are asked to write the definition of the Golden Rule in their own words after reading. As a result of this experiment, the subjects whose state empathy has been promoted through the priming method show significantly higher tax compliance rates.

Herne et al. (2022) investigate how role awareness (the roles of the allocator and receiver), trait empathy, and state empathy affect the allocator's distribution in dictator games. Different from the previous studies reviewed above, they clearly distinguish between state empathy and trait empathy. Since trait empathy is an individual's ability, they measure subjects' trait empathy using the Interpersonal Reactivity Index (IRI) and the Questionnaire of Cognitive and Affective Empathy (QCAE) before the experiment. In addition, they present the following to promote state empathy before making the distribution decision in the empathy-induction condition: "Before you decide how much you will send to the receiver, evaluate how the receiver will feel about receiving different sums of money. Write down your evaluations of the receiver's feelings." For role awareness, they set a role-certainty condition, in which a role is assigned before decision-making, and a role-uncertainty condition, in which after all subjects have made a distribution decision as an allocator, a role is assigned and the payoff is determined by the decision of the subject who is selected as the allocator. As a result, empathy induction and role awareness increase the allocators' average distribution to receivers, but the results of a regression analysis including all factors, role awareness, trait empathy, and state empathy, show that role awareness and some subscales of trait empathy have significant effects on

[2]The Golden Rule is the principle of treating others as one wants to be treated.

the allocator's distribution decision while state empathy does not. Among the subscales of trait empathy, empathy concerns in the IRI and affective empathy (the ability to feel the emotions of others as if they are one's own) in the QCAE have significant influences on the distribution of allocators.

The results of Herne et al. (2022) seem to argue that the promotion of state empathy is not related to altruistic behavior; however, the role uncertainty condition could induce state empathy. In situations where subjects do not know whether they will be the allocator or the receiver, the distribution decision takes into account both situations, in which they would be in two roles. Therefore, in situations where roles are uncertain, people tend to maximize efficiency so that they can accept the distribution regardless of what role is assigned rather than maximizing their payoffs as if they were allocators (Charness & Grosskopf, 2001; Engelmann & Strobel, 2004; Iriberri & Rey-Biel, 2011). Iriberri and Rey-Biel (2011), Engelmann and Strobel (2004), and Charness and Grosskopf (2001), who integrate role uncertainty into their dictator game experiments, obtain similar results showing that role uncertainty decreases selfish behaviors and increases altruistic behaviors. Therefore, the results of Herne et al. (2022) showing that empathy induction has no significant effect on the distribution of allocators could be caused by a failure to consider the effect of state empathy, which is included in the role uncertainty condition. In other words, the results of Herne et al. (2022) also imply the possibility that promoting state empathy has a positive effect on fairness considerations. Furthermore, they suggest that the role uncertainty condition could be a useful treatment to promote state empathy by allowing subjects to consider both the situation of the allocator and the situation of receivers rather than making them imagine others' positions by presenting sentences or phrases to induce state empathy.

Andreoni and Rao (2011) use the role uncertainty condition in the empathy treatment to promote state empathy. The main focus of their study is on how the messages between allocators and receivers affect distribution decisions. The messages are not binding to each other; that is, they are "cheap talk" (Farrell & Rabin, 1996). They compare four types of communication conditions in Experiment 1. 1) The allocator sends the distribution decision and explanation messages to the receiver unilaterally (Explain condition, E). 2) The receiver sends the ask for the distribution and messages to the allocator, and then the allocator reads the receiver's ask and makes a decision to distribute (Ask condition, A). 3) First, the allocator plays the Explain condition, and then the receiver reads it and plays the Ask condition. Finally, the allocator makes the final distribution decision after reading the receiver's ask (Explain and Ask condition, EA). 4) First, the receiver plays the A condition, and then the allocator plays the E condition after reading the receiver's message (Ask and Explain condition, AE). As a result, the share to the receiver significantly increases under the three conditions with an ask, namely, the A, EA, and AE conditions.

They conduct Experiment 2 to add a stage for promoting empathy to Experiment 1. In Stage 1 of Experiment 2, subjects face both the E and A conditions without role assignment, which promotes state empathy. In Stage 2, roles are assigned to each subject, and the E and A conditions of Experiment 1 are conducted. The share to the receiver significantly increases in the E condition of Experiment 2 (E(e) condition) compared to the E condition in Experiment 1. When comparing the A condition of Experiment 2 (A(e) condition) to the A condition in Experiment 1, the share to the receiver slightly increases, but the difference is not significant. Moreover, the share to the receiver in the A(e) condition is similar to those in the EA and AE conditions with bilateral messages. The study shows that promoting state empathy has effects similar to those of bilateral

communication.

These previous studies show that state empathy has a significant effect increasing the fairness of distribution in dictator games. In particular, as mentioned above, the role uncertainty condition effectively promotes state empathy by allowing people to experience others' situations. Therefore, in our experiments, I adopt the A(e) condition of Andreoni and Rao (2011); that is, I promote state empathy through role uncertainty. Moreover, I verify whether there is a difference in distribution behavior between the empathy treatment, which includes a stage wherein state empathy is promoted through role uncertainty, and a traditional dictator game among people with plural fairness ideals. I give the details in Chapter 4.

Chapter 2

Heterogeneity in the Motivations for Punishment

2.1 Introduction

People punish free-riding and other undesirable behaviors. People sometimes still tend to punish even if there are no future benefits, only costs. For example, in a public goods game with a one-shot interaction or with iterations over stranger matching, punishments do not provide any direct benefits to those subjects who punish others because they will not meet the target again and cannot benefit from the target's reaction to punishment—which, in most cases, is cooperation. However, punishment of free riders, which also encourages cooperation, is frequently observed (Balliet, Mulder, & Van Lange, 2011). What, then, makes people punish others even when it is hard to expect any future benefit? Unfortunately, few studies have concentrated on the motivations for punishment, and most of them are focused on punishment's effects (see Balliet et al. (2011)).

This chapter focuses on the motivation for imposing costly punishments in social dilemma situations. When someone free rides, the action has two undesired components. One is the betrayal behavior, which is unkind and selfish, and the other is the inequality between the free rider's higher payoff and the cooperator's lower payoff. Punishment is a reaction to undesired things, and there are two streams of research about the motivations behind punishment. The first motivation is reciprocity, which leads to retaliation against undesired behaviors, and the second is inequality aversion, which reduces undesired inequality. These two motivations have been studied for a long time; however, the streams of research on costly punishments in social dilemmas consider merely whether two or more motivations are able to coexist within an individual, and only a few studies have considered both motivations concurrently.

However, by conducting 10 simple games, Leibbrandt and López-Pérez (2012) show that four motivations—selfishness, inequality aversion, spite, and reciprocity—could coexist. In addition, Xiao and Bicchieri (2010) show the coexistence of three motivations—payoff maximization, inequality aversion, and reciprocity—in trust game experiments. These results imply that there could be heterogeneity in the motivations to punish.

We believe that there is also heterogeneity in the motivation to punish in public goods games, and we try to classify subjects according to their motivations. In our experiment, to classify subjects according to their motivation for punishment and to verify the characteristics of the subjects exhibiting each motivation, we conduct modified public goods games with costly punishments. Since reciprocity is based on others' behaviors and inequality aversion on unequal payoffs, we use an uncertainty factor as noise, which weakens

the link between behaviors and payoffs. This noise is randomly and individually added to the payoffs in a public goods game after the contribution decisions are made. For example, when negative noise is added to a free rider's payoffs in a public goods game, it decreases the free rider's high payoffs. Conversely, when positive noise is added to the cooperator's payoffs, it increases the cooperator's low payoffs. Thus, people can choose who should be punished: free riders, high earners, or both. Based on the individuals' punishment patterns in these games, we estimate and classify the subjects' types. Then, we analyze each type's behavioral pattern in more detail.

In our results, the shares of the reciprocal type are the highest, but a significant percentage of people punish solely based on inequality aversion. As a robustness test, we conduct random income games (Dawes et al., 2007)[1] and explore whether intentions matter (Rabin, 1993; Levine, 1998) only to the reciprocal type. As expected, subjects classified as being of the reciprocal type have the greatest decrease in punishments. This result clarifies our classification by indicating the difference between the reciprocal type and the inequality-averse type when there is no room for reciprocity. Moreover, we find that the reciprocal type strongly punishes free-riding behaviors but also imposes some punishment for payoff inequality.

This chapter proceeds as follows. Section 2.2 explains our experiments, which are designed to distinguish the motivations for punishment. Section 2.3 provides the experimental results, and Section 2.4 gives a summary of this study and a discussion.

2.2 The Experiments

Our experiments are public goods games with costly punishments (Fehr & Gächter, 2000). Each session is composed of two games. All subjects participate in both games in order, and we call the first game the baseline game and the second game the random income game (Dawes et al., 2007). After seating all participants, we distribute the instructions for the baseline game only because the random income game starts without warning. After the computer reads the instructions aloud, we provide a problem set to check the participants' understanding before starting the baseline game. When every participant solves each problem correctly, the games begin.

At the beginning of each round in the baseline game, all participants are endowed with 20 tokens and divided into new four-member groups. Since the group composition changes randomly every round, there is little room to consider the future benefits from punishing others. Each round consists of two stages: the contribution stage (Stage 1) and the punishment stage (Stage 2). In Stage 1, participants decide the amount x_i to contribute to a public account, where $0 \leq x_i \leq 20$ and the marginal per capita return on the public account is 0.5. To distinguish between the motivations for punishment, we add the uncertainty factor ε_i to the Stage 1 payoff function. Hence, i's payoff at the end of Stage 1 is given by:

$$\pi_i = 20 - x_i + 0.5 \sum_{j=1}^{4} x_j + \varepsilon_i. \tag{2.1}$$

ε_i is an integer in $[-8, 8]$ and follows a random sequence with mean 0 over all rounds[2].

[1] The procedure for the random income game in our experiments is explained in Section 2.2.

[2] Generally, the average contribution in a linear public goods game as a percentage of the total endowment is 37.7 (Zelmer, 2003). Based on this fact, we set ε_i to range between -8 and 8, which is 40% of our endowment, 20. If one's payoffs are calculated with ε_i equal to 8, the effect of having two free-riders is the same as having everyone contribute at the average level.

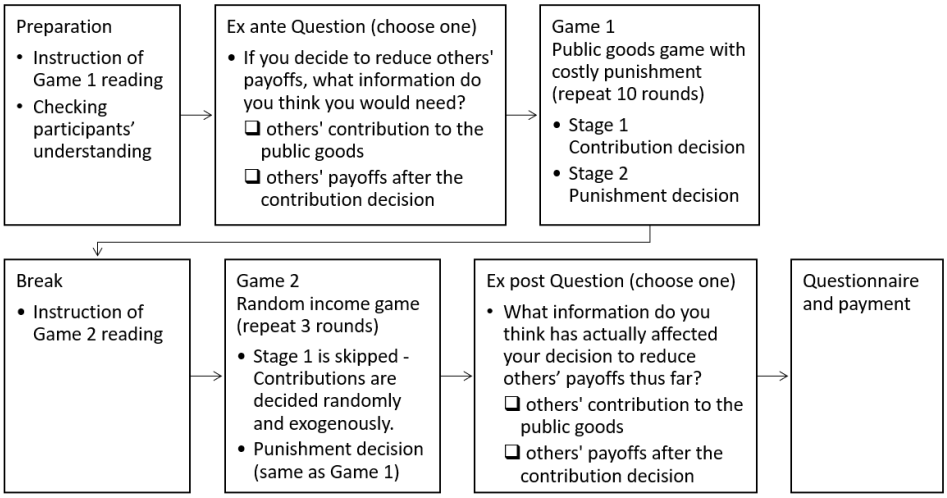


Figure 2.1: Experimental Procedure.

The sequence of ε_i follows a uniform distribution, and the subjects do not know this distribution. However, they do know that ε_i is assigned randomly. This uncertainty factor weakens the causal relation between behaviors and payoffs. For example, if someone contributes ten tokens but is assigned $\varepsilon_i = 8$, his or her payoff is the same as if he or she had contributed two tokens with $\varepsilon_i = 0$. If someone contributes 0 tokens and is assigned $\varepsilon_i = -8$, this is the same as contributing eight tokens with $\varepsilon_i = 0$. That is, cooperators may not receive a lower payoff than free-riders, and free-riding behaviors do not necessarily generate higher payoffs. In this situation, when subjects want to punish others, they could choose their criterion, behavior or payoff. This allows us to classify the subjects according to their motivations. All participants have their own sequence of ε_i, so each group member may be assigned a different ε_i.

When all participants decide their contributions to the public good, they move on to Stage 2 and receive nine additional tokens that can be used to inflict punishments.[3] In this stage, the participants are informed of the other three members' contributions and payoffs from the first stage. Of course, participants know their own contributions and payoffs. They can punish the other members by using their tokens—up to three for each member—and each token decreases the target's payoff by three tokens. Hence, subject i's final payoff for each round is as follows:

$$\pi_i = 20 - x_i + 0.5\sum_{j=1}^{4} x_j + \varepsilon_i + 9 - \sum_{j=1}^{4} p_{ij} - 3\sum_{j=1}^{4} p_{ji}. \tag{2.2}$$

The baseline game has ten rounds, and participants know these payoff functions and processes, including the number of rounds.

After finishing the tenth round of the baseline game, we ask participants to play an additional game, a random income game (Dawes et al., 2007). In the random income game, the contribution stage is skipped. The reason why the contribution decision is eliminated is to clarify the subjects' motivations as follows: people who punish others for their contributions will stop punishing in a random income game because there are

[3]The reason for the nine additional tokens for punishment is to reduce the impact of the payoff at the contribution stage on punishment behavior, namely the income effect.

Table 2.1: Summary of Sessions.

Session number	Treatment	Total number of groups	Total number of subjects
1	Baseline (10 rounds) + Random income game (1 round)	7	28
2–5	Baseline (10 rounds) + Random income game (3 round)	27	108
Total		34	136

no intentional behaviors, and those who punish others for their payoffs will continue to impose punishments. The random income game is played for a total of three rounds[4], although participants are not aware of the total number of rounds that will be played in order to avoid end-game effects.[5] The random income game is the same as the baseline game except during Stage 1. In Stage 1 of the random income game, contributions are decided randomly and exogenously, and all participants move directly to Stage 2. In other words, each participant makes only a punishment decision based on the others' computer-determined contributions. For a more rigorous comparison of the two treatments, we use the data from three rounds in the baseline game. More specifically, in the first round of the random income game, participants make punishment decisions based on information of the data of group compositions and member contributions from the third round of the baseline game as the others' computer-determined contributions. In the second round of the random income game, they do it based on those from the seventh round of the baseline game, and in the third round of the random income game, they do it based on those from the fifth round of the baseline game. However, the participants are not informed of this. We distribute new instruction after finishing the baseline game and the computer reads it aloud. Through this random income game, our classification of motivation types based on the punishment behaviors in the baseline games will be verified.

Additionally, both before starting the baseline game and after finishing the random income game, we ask the following ex ante and ex post questions: "If you decide to reduce others' payoffs, what information do you think you would need?" and "What information do you think has actually affected your decision to reduce others' payoffs thus far?". Participants are asked to choose one of the following two options: 1) others' contribution to the public goods and 2) others' payoffs after the contribution decision has been made. From the answers to these questions, we can make conjectures about the relation between revealed motivations and stated motivations. Figure 2.1 briefly shows the overall experimental procedure.

Finally, to identify the characteristics of each type, we distribute a questionnaire. We extracted some questions from the National Opinion Research Center's General Social Survey and the Japanese Temperament and Character Inventory (Kijima, 1996).

The experiments were conducted at Waseda University during Summer and Fall 2018. We ran five sessions with 24 to 28 subjects in each session, and all 136 participants were students from Waseda University (Table 2.1). Participants' majors vary such as economics,

[4]We conduct one round of random income game in the first session to check the validity as a robustness test. Since the random income games are played after finishing baseline game, number of rounds of random income game do not affect our main results, that is, subjects' types and their behaviors.

[5]Since the random income game has a similar structure to the baseline game, the learning effect may be strong if ten rounds are played. Moreover, there is also the length of the experiment, so the random income game is shortened to three rounds. Subjects are not informed how many rounds would be played because of the possibility of end-game effects when the length of the game is shortened.

education, humanities, law, commerce, politics, engineering, etc. The experiments were computerized with z-Tree (Fischbacher, 2007). Each session lasted 60 minutes, and the show-up payment was ¥700 (≈ \$6.2), and average earnings were ¥1890 (≈ \$16.6)[6].

2.3 Results

In this section, we first estimate each subject's type and classify the players. Then, we examine the differences between types in terms of behaviors and characteristics.

2.3.1 Estimation of Subject Types

As discussed above, the uncertainty factor ε_i weakens the correlation between contributions and payoffs.[7] In this situation, subjects face the problem of what to punish: others' behaviors or their payoffs. Thus, we hypothesize that subject types can be distinguished based on the motivations for punishment as follows:

1) *The self-interested type* (Type S) does not punish at all because punishment does not generate any benefits, only costs.
2) *The reciprocal type* (Type R) chooses to punish others' unkind behaviors. This type focuses on others' contribution levels and punishes low contributors or free riders.[8]
3) *The inequality-averse type* (Type IA) chooses to punish those who earn higher payoffs. This type wants to reduce the payoff gap by using punishments.
4) *The "other" type* (Type O) punishes others inconsistently. This type could be motivated by social preferences other than reciprocity and inequality aversion or could inflict punishments for accidental or unknown reasons.

Type S chooses a dominant strategy that maximizes his or her monetary payoffs. Type R is suggested by models of fairness intentions and kindness (Rabin, 1993; Levine, 1998; Dufwenberg & Kirchsteiger, 2004; Cox et al., 2007), and Type IA is suggested by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), as reviewed above. Type O players could be motivated by other social preferences, such as spite (Leibbrandt & López-Pérez, 2012; Hilbe & Traulsen, 2012), envy (Kirchsteiger, 1994), and preference for competition (Bault et al., 2008). Alternatively, Type O players may inflict punishments inconsistently for accidental or unknown reasons. As our study lacks tools to identify the specific motivations behind this inconsistency in punishments, we refer to this group's players as the "other" type.

Carpenter and Matthews (2012) shows that subjects punish based on two norms: 1) a comparison of the individual contributions of the subject himself or herself and the target and 2) a comparison of the group average contribution and that of the target. Thus, we consider two regression models corresponding to individual and group comparisons.

[6] We convert Japanese yen into US dollars at the exchange rate of \$1=¥113, which was the average rate at the time that the experiments were conducted, Fall 2018.

[7] The results of a correlation test show that there is a weak correlation between subject i's contributions and payoffs ($r = -0.2998$, 1% significance level). Moreover, subject j's contribution and subject i's payoff are not highly correlated ($r = 0.4078$, 1% significance level).

[8] We use the concept of *negative reciprocity* in Fehr and Gächter (2000): "in the case of reciprocity, the actor is responding to friendly or hostile actions even if no material gains can be expected".

To classify subject types based on this hypothesis, we establish the following two regression models:

$$p_{ij} = \beta_1 + \beta_2 \min\{Con_j - Con_i, 0\} + \beta_3 \max\{Pay_j - Pay_i, 0\} \quad \text{(M1)}$$

$$p_{ij} = \beta_4 + \beta_5 \min\{Con_j - GCon_{-j}, 0\} + \beta_6 \max\{Pay_j - GPay_{-j}, 0\} \quad \text{(M2)}$$

where p_{ij} on the left-hand side is the number of punishment tokens that i spends on target j. In M1, Con_j is target j's contribution, Con_i is i's own contribution, Pay_j is target j's payoff in Stage 1, and Pay_i is i's own payoff in Stage 1. In M2, $GCon_{-j}$ is the group average contribution excluding that of target j, and $GPay_{-j}$ is the group average payoff excluding that of target j. Because of multicollinearity, we separate the variables in these two models according to two criteria: deviations from one's own contribution/payoff and deviations from the group average.

The reason why we use the min/max function in these models is the uncertainty factor, which might cause cooperators' payoffs to be higher than free riders'. In this case, a subject who is motivated by inequality aversion will punish a cooperator because of the payoff inequality among the members in his or her group. This punishment decision seems to be an antisocial punishment, which means that it punishes socially friendly behaviors and cooperation.[9] However, in our experiments, such punishments may not be directed against cooperation but rather may be driven by an aversion to inequality. To rule out the possibility of such misinterpretation, we consider only negative gaps in contributions and disadvantageous inequality in payoffs.[10]

We estimate the subject's type with linear regressions. Before estimation, we separate the subjects who had $p_{ij} = 0$ throughout the baseline game and classify them as Type S ($N = 52$). Then, regressions for the remaining participants who punished others at least once are estimated separately for each participant. Each subject participated in ten rounds in the baseline game, and in each round, they made decisions about punishments for the other three members. That is, there are 30 observations of punishment decisions p_{ij} for each subject, and we can perform regression analyses with p_{ij} as the dependent variable separately for each subject.

We determine the types of subjects through the following process: 1) We select coefficients with high significance levels with the lowest p value in both Models M1 and M2. 2) When two or more coefficients have the same significance level of $p = 0.0000$ in both models, we compare their magnitudes. We normalize[11] all variables ($\min\{Con_j -$

[9] To address the possibility of antisocial punishment, new variables ($\max\{Con_j - Con_i, 0\}$, $\max\{Con_j - GCon_{-j}, 0\}$) are added to the regression Models M1 and M2 above and analyzed for each subject. We find that 7.4% of all participants (10 people) seem to inflict both prosocial and antisocial punishment at the same time. However, this percentage is very low compared to the results in previous studies (15% in Fehr and Gächter (2002) and 22% in Gächter and Herrmann (2009) as calculated by Thöni (2014)), and the significance and the magnitude of the coefficients on the antisocial variables are also lower than those on the prosocial variables. Therefore, in this study, we do not classify subjects who impose antisocial punishments as a distinct type. See Appendix B for more details.

[10] When we do not use the min/max function to estimate the subjects' types, the types of a few subjects change, and the overall significance of their coefficients decreases.

[11] We use min-max normalization. For example, in the case of $\min\{Con_j - Con_i, 0\}$, we calculate $\text{norm\_}\min\{Con_j - Con_i, 0\}$ as follows:

$$\begin{aligned}\text{norm\_}\min\{Con_j - Con_i, 0\} &= \frac{\min\{Con_j - Con_i, 0\} - \min\limits_{\substack{k \neq l \\ k,l \in I}} \min\{Con_l - Con_k, 0\}}{\max\limits_{\substack{k \neq l \\ k,l \in I}} \min\{Con_l - Con_k, 0\} - \min\limits_{\substack{k \neq l \\ k,l \in I}} \min\{Con_l - Con_k, 0\}} \\ &= \frac{\min\{Con_j - Con_i, 0\} - (-20)}{0 - (-20)} = \frac{\min\{Con_j - Con_i, 0\} + 20}{20},\end{aligned}$$

where I is the set of subjects.

Table 2.2: Type Classification.

Type	Number of subjects	Percentage
Self-interested (Type S)	52	38%
Reciprocal (Type R)	40	29%
Inequality Averse (Type IA)	35	26%
Other (Type O)	9	7%
Total	136	100%

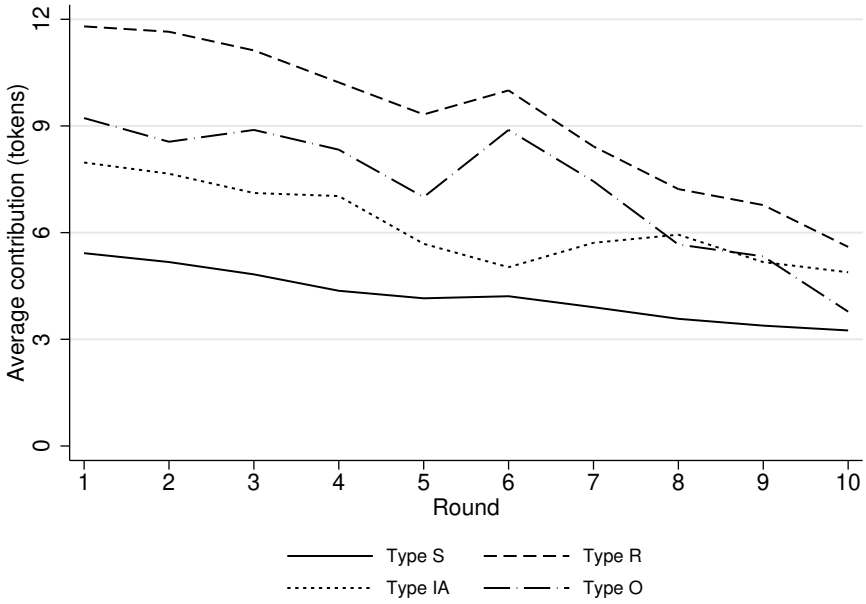


Figure 2.2: Cooperative Behaviors by Type.

$Con_i, 0\}$, $\min\{Con_j - GCon_{-j}, 0\}$, $\max\{Pay_j - Pay_i, 0\}$ and $\max\{Pay_j - GPay_{-j}, 0\}$) to a range of 0 to 1 before comparing the magnitudes of the coefficients.[12]

Subjects are classified as Type R if they have either β_2 from Model M1 or β_5 from Model M2 selected, or both. Similarly, subjects with either β_3 from Model M1 or β_6 from Model M2 selected, or both, are classified as Type IA. Based on this process, some subjects have selected coefficients in both models (individual comparison, M1, and group comparison, M2), while others have selected coefficients in only one of the models. As we mentioned above, our two models represent two different criteria or norms (Carpenter & Matthews, 2012), not different motivations. In addition, we have three subjects for which both the Type R and Type IA coefficients have the same p value of 0.0000, so we compare the magnitudes of their significant coefficients and assign their type based on the largest coefficient, following step 2) in the process described in the previous paragraph. Finally, subjects who do not have any significant coefficients are classified as Type O. The results of the type classification are shown in Table 2.2.

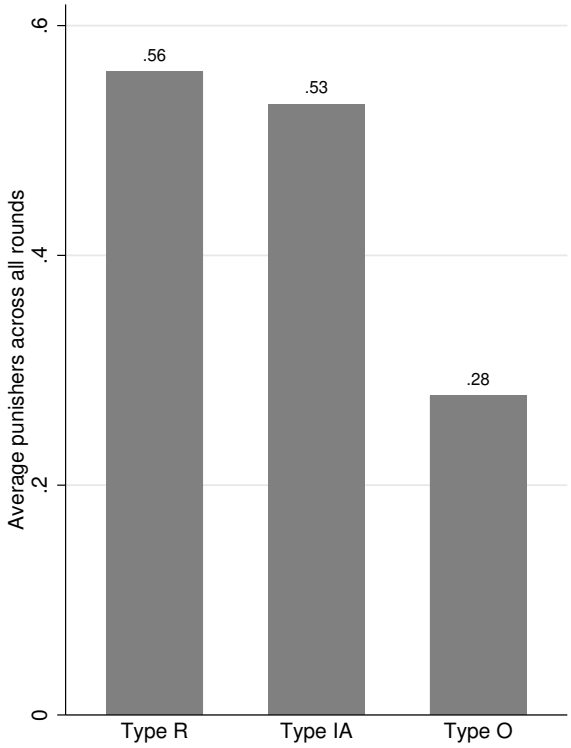


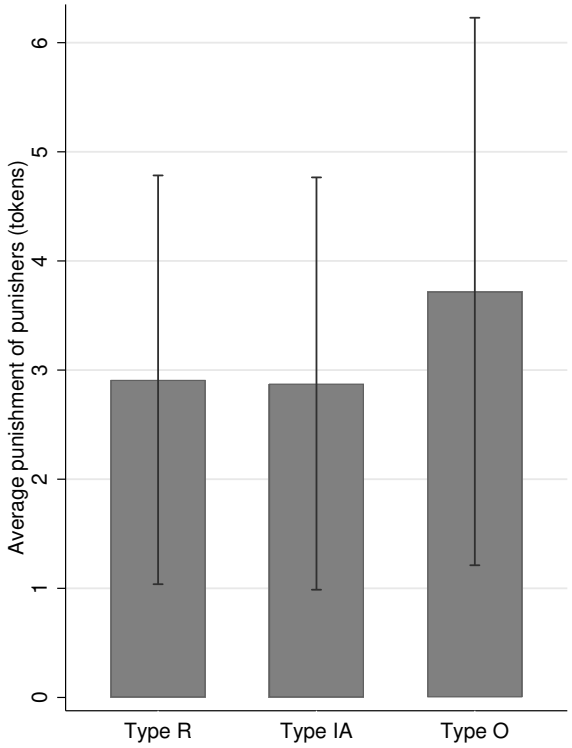
Figure 2.3: Proportion of Punishers.

2.3.2 Cooperation and Punishment Behaviors

Figure 2.2 shows the average contributions by round for each type. The type specific average contributions across all rounds are 9.2 tokens for the Type R subjects (Type R for brevity), 6.2 tokens for the Type IA subjects (Type IA for brevity), 7.3 tokens for the Type O subjects (Type O for brevity), and 4.2 tokens for the Type S subjects (Type S for brevity). The average level of contributions tends to decrease over time, which is a standard result in public goods games. Type R is the most cooperative, and Type S is the least cooperative, as shown in Figure 2.2. More specifically, in round 1, Type S subjects contribute significantly less than any other type (t-test; compared with Type R: $p = 0.0000$, with Type IA: $p = 0.0326$, with O: $p = 0.0661$), and Type R subjects contribute significantly more than any other type except Type O (t-test; compared with Type IA: $p = 0.0026$, with Type O: $p = 0.1391$). In round 10, Type S subjects still contribute significantly less than any other type except Type O (t-test; compared with Type R: $p = 0.0157$, with Type IA: $p = 0.0463$, with Type O: $p = 0.3826$). However, Type R subjects contribute significantly more than Type S subjects (t-test; compared with Type IA: $p = 0.2390$, with Type O: $p = 0.1587$). In other words, among all types, Type R shows the largest declines in contributions.

Punishment behaviors by type are presented in Figures 2.3 and 2.4. Figure 2.3 depicts the average proportion of punishers who use a positive number of tokens for punishment across all rounds, and Figure 2.4 presents the average number of punishment tokens used by punishers by type. Specifically, the proportion of Type O punishers is significantly

[12] We observe that the coefficients with the lowest p values are mostly the largest, or at least not significantly different from the values of the other significant coefficients.



Note: Bars depict standard deviations.

Figure 2.4: Average Punishment of Punisher.

lower than that of Type R and Type IA punishers at the 1% level (Wilcoxon rank-sum test, $p = 0.0000$ and $p = 0.0000$, respectively). Additionally, the proportions of Type R and Type IA punishers are not significantly different (Wilcoxon rank-sum test, $p = 0.3575$). However, on average, each punisher spends approximately two tokens on each punishment.[13]

These results on cooperation and punishment behaviors provide further evidence for our classification of types. Individuals classified as Type R, who punish others' undesired behaviors, appear to focus on others' behavior because they behave cooperatively. On the other hand, Type S subjects, who do not punish at all, are the most uncooperative. In other words, they consistently refrain from making contributions or punishing. Their pattern of uncooperative behavior, coupled with the cooperative behavior of Type R, lends support to the plausibility of our classification. Furthermore, in the case of Type O individuals, who punish others inconsistently, the proportion of punishers is significantly lower than that among other types. This may indicate that social preferences other than reciprocity and inequality aversion that serve as motivations for punishment are negligible or that this type's punishments are accidental.

2.3.3 Revealed Motivation vs. Stated Motivation

As mentioned in Section 2.2, we asked the following ex ante and ex post questions: "If you decide to reduce others' payoff, what information do you think you will need?" and "What information do you think has actually affected your decision to reduce others' payoffs thus

[13]See Appendix C for a more detailed analysis of punishment behaviors.

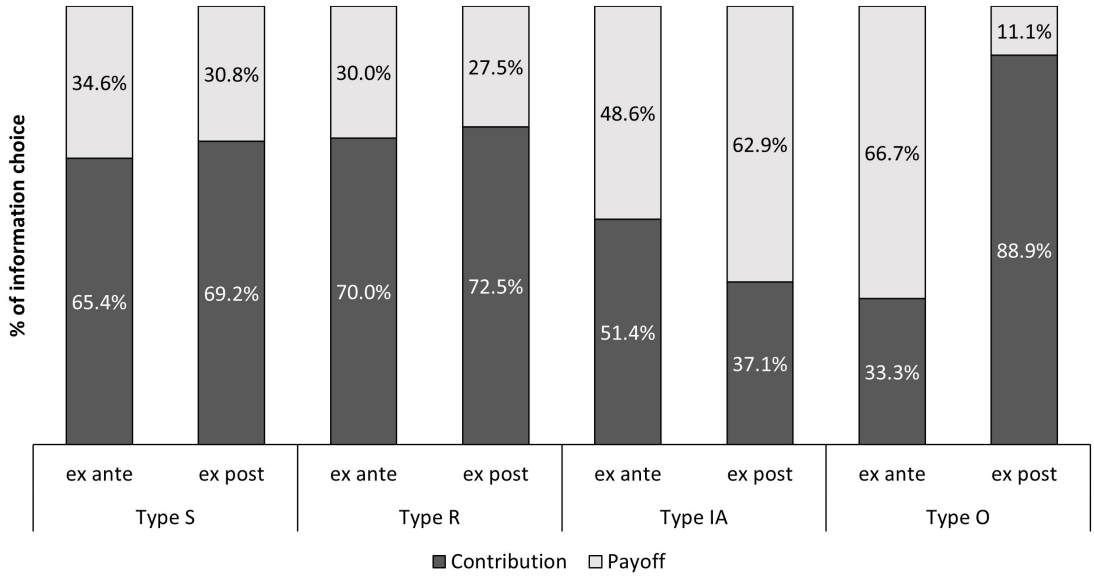


Figure 2.5: Information Choice.

far?".[14] Because there are only two options, the other's contributions and the other's payoffs in stage 1, the responses are presented in Figure 2.5.

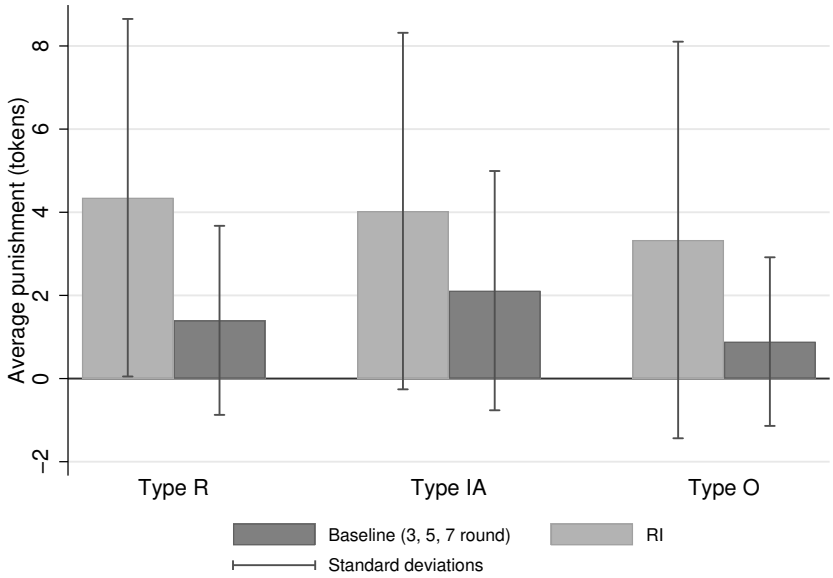
Generally, subjects choose the information that they need according to their type, as shown in Figure 2.5. Type S never punishes others, but in terms of information, they want to know others' contribution levels. In the case of Type IA, 14.3% of such subjects change their mind from others' contributions to others' payoffs as the information necessary to make punishment decisions. Some of the Type IA subjects at first believe themselves to be Type R; however, as the game progresses, they eventually realize that they have been motivated to punish by the inequality in payoffs. Type O subjects change their choice the most, and this result implies that most Type O subjects do not punish consistently and may punish without a clear motivation. This is one of the reasons why they are classified as Type O and do not have any significant coefficients.

These results regarding information choices suggest that people are generally aware of their motivations for punishment. Moreover, people who are not willing to impose a costly punishment without a future benefit (Type S) prefer reciprocity as a motivation, and people who are averse to inequality (Type IA) often realize their motivation only later. However, 27.5% of Type R subjects (ex post information choice in the second panel of Figure 2.5) and 37.1% of Type IA subjects (ex post information choice in the third panel of Figure 2.5) still misunderstand their own motivation to punish at the end of the experiment.

2.3.4 Comparison with the Random Income Game

To clarify the results of our type estimation, we now turn to the results of the random income game. Since the random income game does not include a contribution stage and the computer randomly decides each member's contributions, there is no intentionality behind the contributions made. Thus, we expect that only Type R subjects, who punish

[14] While the possibility exists for subjects to self-control their behavior by answering ex ante question, the results in the figure show that around 30% of subjects either do not control their punishment behavior or misunderstand it. Thus, the effect of self-control appears to be negligible.



Note: Bars depict standard deviations.

Figure 2.6: Average Total Punishment in the the Third, Fifth, and Seventh Rounds of the Baseline Game and the Random Income Game.

others' contribution decisions, will not punish, in line with the kindness function of Rabin (1993) and the model of altruism in Levine (1998). In the cases of Type IA, we expect that these subjects will punish others who obtain higher payoffs because inequality aversion is their motivation for punishment (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000).

In our experiments, as described in Section 2.2, we randomly shuffled the order of the third, fifth, and seventh rounds of the baseline games to obtain data on others' contributions. We did not disclose to subjects the origin of the data or the number of rounds. Subjects were informed that the contributions were determined by a computer and that the game would be repeated several times. This allows us to directly compare the results of the baseline game and the random income game. Figure 2.6 displays the average total punishment in the third, fifth, and seventh rounds of the baseline game and the random income game, while Figure 2.7 shows the average individual differences in total punishment between the baseline game and the random income game. Our results indicate that Type R subjects decrease their punishment the most in the random income game at the 1% level (t-test; $p = 0.0001$), although Type IA and Type O subjects also exhibit significant decreases at the 5% and the 10% levels, respectively (t-test; $p = 0.0159$ and $p = 0.0881$, respectively). However, there are no significant differences in the average individual differences between the baseline game and the random income game (t-test; Type R vs. Type IA: $p = 0.1226$, Type R vs. Type O: $p = 0.3744$, Type IA vs. Type O: $p = 0.3432$).

However, as Figure 2.8 shows, only in the case of Type R is the percentage of zero-punishers who stop punishing others significantly increased (Wilcoxon rank-sum test, $p = 0.0000$) in the random income game. There are no significant differences between the two games in the case of Type IA and Type O (Wilcoxon rank-sum test, $p = 0.3326$ and $p = 0.1587$, respectively). Thus, all subjects tend to reduce their punishment in the random income game, but for Type R specifically, there is a significantly higher tendency to completely stop punishing in this game.

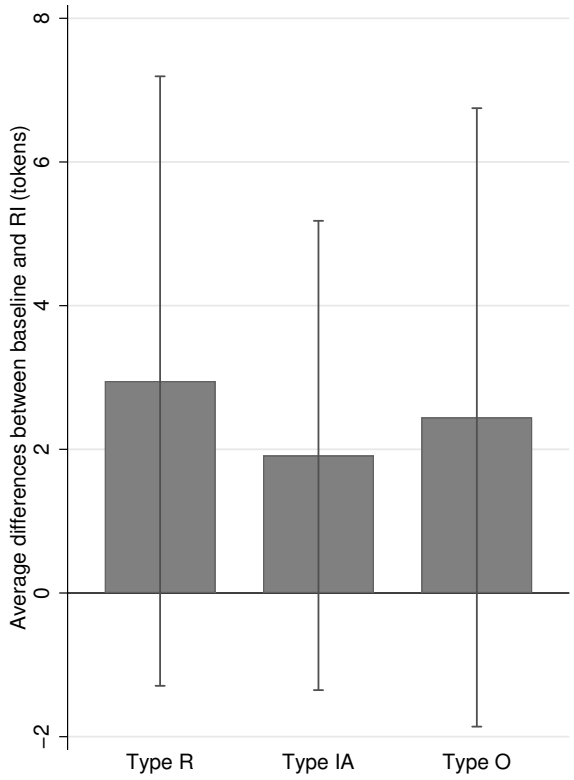


Figure 2.7: Average Difference in Punishment between the Baseline Game and the Random Income Game.

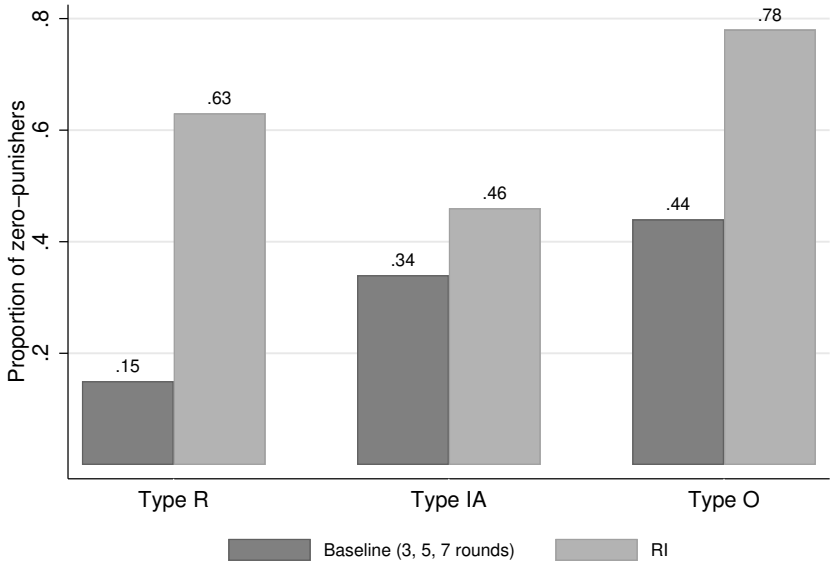


Figure 2.8: Percentage of 0-Punishers.

Table 2.3 presents the results of Tobit regressions[15] that uses the third, fifth, and seventh rounds of the baseline game and all rounds of the random income game. The regression models are as follows:

$$\begin{aligned} p_{ij} =& \beta_1 + \beta_2 Tr + \beta_3 \min\{Con_j - Con_i, 0\} + \beta_4 \max\{Pay_j - Pay_i, 0\} \\ &+ \beta_5 Tr \times \min\{Con_j - Con_i, 0\} + \beta_6 Tr \times \max\{Pay_j - Pay_i, 0\} \quad \text{(M3)} \\ p_{ij} =& \beta_7 + \beta_8 Tr + \beta_9 \min\{Con_j - GCon_{-j}, 0\} + \beta_1 0 \max\{Pay_j - GPay_{-j}, 0\} \\ &+ \beta_1 1 Tr \times \min\{Con_j - GCon_{-j}, 0\} + \beta_1 2 Tr \times \max\{Pay_j - GPay_{-j}, 0\} \quad \text{(M4)} \end{aligned}$$

where Tr is a dummy variable that equals 0 for the baseline game and 1 for the random income game. We include interaction variables between the treatment and the motivations. $Tr \times \min\{Con_j - Con_i, 0\}$ and $Tr \times \min\{Con_j - GCon_{-j}, 0\}$ indicate the interaction between Tr and reciprocity, and $Tr \times \max\{Pay_j - Pay_i, 0\}$ and $Tr \times \max\{Pay_j - GPay_{-j}, 0\}$ indicate the interaction between Tr and inequality aversion. Similarly to how we adopt both Models M1 and M2 for the type classification, we use Models M3 and M4 to indicate the different criteria for punishment under individual comparisons and group comparisons, respectively, as explained in Section 2.3.1.

Since the random income game excludes the contribution decision, we hypothesize the following: (A) only Type R, who punishes others' contribution decisions, will stop punishing others in this game, and (B) Type IA will continue to punish those who obtain higher payoffs because this type is motivated to punish by inequality aversion.

First, regarding hypothesis (A), Type R subjects, who punish others' contribution decisions, considerably reduce their punishment for others' behaviors in the random income game, as there are no contribution decisions. This observation is supported by the fact that the coefficients on both the interaction variables, $Tr \times \min\{Con_j - Con_i, 0\}$ and $Tr \times \min\{Con_j - GCon_{-j}, 0\}$, are significant and that their magnitudes (0.2473 and 0.2508) offset the coefficients on reciprocal punishment by over 50%: -0.3971 for $\min\{Con_j - Con_i, 0\}$ and -0.5144 for $\min\{Con_j - GCon_{-j}, 0\}$.

Second, regarding hypothesis (B), while the significant coefficients on Tr in Models M3 and M4 suggests that participating in the random income game can lower Type IA's punishment levels, this effect only offsets the constant. However, the coefficients on the interaction variables, $Tr \times \max\{Pay_j - Pay_i, 0\}$ and $Tr \times \max\{Pay_j - GPay_{-j}, 0\}$, are not significant in either model. This finding supports hypothesis (B), as it indicates that Type IA subjects maintain their punishment for payoff inequality regardless of others' intentions.

Interestingly, Type R subjects exhibit a tendency to punish both free-riding behavior and payoff inequality. The coefficient on inequality aversion, represented by $\max\{Pay_j - Pay_i, 0\}$, is found to be statistically significant at the 5% levels.[16] This finding suggests that individuals who punish based on reciprocity also consider inequality aversion, unlike those who punish based only on inequality aversion. Previous studies on costly punishment have predominantly focused on reciprocity, overlooking the motivation of inequality aversion. Thus, it is crucial to recognize the importance of inequality aversion along with reciprocity as a driving force for inflicting punishment.

[15] We also conduct a fixed effects panel analysis, but the results are essentially similar. For consistency within this thesis, we present the results of the Tobit regression analysis.

[16] In fact, this result seems to be inconsistent with our hypothesis for Type R. Thus, to check the validity of our type estimates, we added an interaction term to M1 and M2. We observe that 35% of Type R subjects and 31% of Type IA subjects also have a significant interaction term with their main motivation. However, when we perform the same regression as in Table 2.3 with only those subjects who do not have a significant interaction term, the significance of $\max\{Pay_j - Pay_i, 0\}$ in Type R is still significant at the 5% level.

Table 2.3: Results of Tobit Regressions.

Dependent variable: Punishment level	Type R	Type IA	Type O
M3 Individual comparison			
Tr (0: baseline, 1: RI)	-0.5229 (0.6792)	-1.0792^{*} (0.4493)	-6.0141^{*} (2.7622)
$\min\{Con_j - Con_i, 0\}$	-0.3971^{***} (0.0689)	-0.0242 (0.0624)	0.2639 (0.2465)
$\max\{Pay_j - Pay_i, 0\}$	0.1208^{*} (0.0515)	0.1886^{***} (0.0455)	0.1743 (0.1474)
$Tr \times \min\{Con_j - Con_i, 0\}$	0.2473^{**} (0.0942)	0.0979 (0.0972)	0.9972 (0.7561)
$Tr \times \max\{Pay_j - Pay_i, 0\}$	-0.0399 (0.0763)	0.0314 (0.0669)	0.6565 (0.3826)
Constant	-4.8546^{***} (0.6890)	-2.2902^{***} (0.3748)	-3.7698^{*} (1.4502)
M4 Group comparison			
Tr (0: baseline, 1: RI)	-1.5858^{*} (0.6687)	-1.4504^{**} (0.4474)	-4.4788^{*} (2.1286)
$\min\{Con_j - GCon_{-j}, 0\}$	-0.5144^{***} (0.0827)	-0.0099 (0.0715)	0.1431 (0.3218)
$\max\{Pay_j - GPay_{-j}, 0\}$	0.1054 (0.0556)	0.2802^{***} (0.0502)	0.1813 (0.2038)
$Tr \times \min\{Con_j - GCon_{-j}, 0\}$	0.2508^{*} (0.1143)	-0.0512 (0.1078)	-0.1563 (0.5429)
$Tr \times \max\{Pay_j - GPay_{-j}, 0\}$	0.0401 (0.0853)	0.0096 (0.0700)	0.1067 (0.3360)
Constant	-4.2329^{***} (0.6016)	-2.3133^{***} (0.3545)	-4.3030^{**} (1.6100)
Number of observation	588	546	150

Note: Standard errors are given in parentheses. We use the data from the 3rd, 5th, and 7th rounds in the case of baseline game. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

In summary, the results of the random income game provide support for our classification of subjects, as shown in Table 2.2. Specifically, individuals who punish based on reciprocity (Type R) reduce their punishment significantly when there is no intention behind others' behaviors. On the other hand, those who punish for payoff inequality (Type IA) also tend to decrease their punishment, but they maintain their punishment for payoff inequality regardless of others' intentions.

2.3.5 Differences by Gender and Social Characteristics

We now move on to an analysis of the characteristics of each type based on the questionnaire. After the last round of the random income game, the questionnaire was provided to all participants at the same time. It contains questions about basic personal information (gender, year in school, major, etc.) and questions taken from the National Opinion Research Center's General Social Survey[17] and the Japanese Temperament and Character

[17]We choose three questions to measure trust: 1) GSS Trust: "Generally, would you say that most people can be trusted or that you cannot be too careful in dealing with people?" (1: "most people can be trusted", 0: "cannot be too careful"); 2) GSS Fair: "Do you think most people would try to take advantage

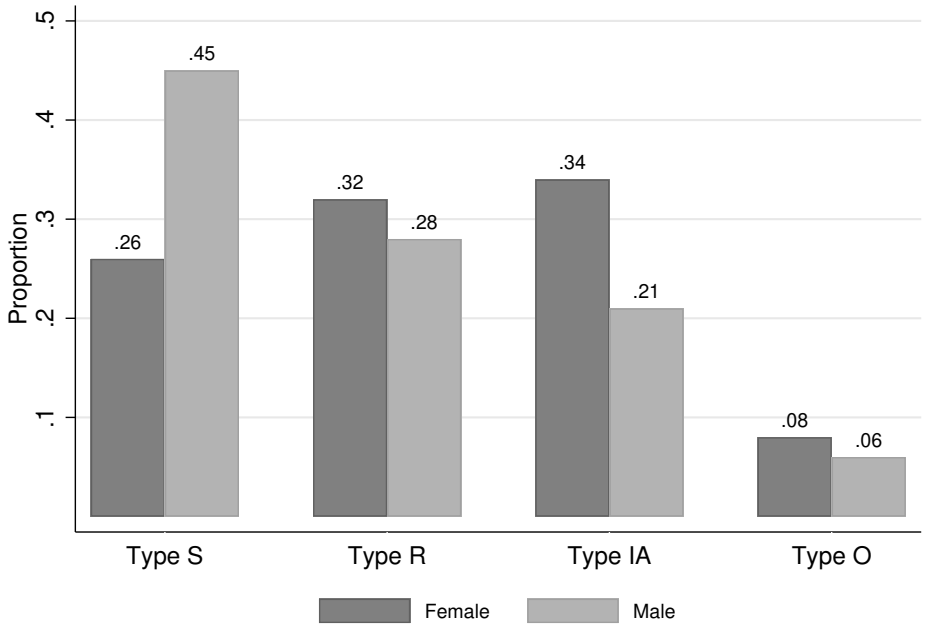


Figure 2.9: Gender Differences.

Inventory (Kijima, 1996), which is used as a personality test. Specifically, we extract five questions from each of the five subscales of the Cooperativeness (CO)[18] dimension of the Temperament and Character Inventory (TCI) (social acceptance vs. social intolerance; empathy vs. social disinterest; helpfulness vs. unhelpfulness; compassion vs. vengefulness; pure-hearted conscience vs. self-serving advantage). From the questionnaire, we obtain two significant results.

First, there are gender differences among the types. Figure 2.9 describes the gender distribution by type. We have data from 86 male subjects and 50 female subjects. Types S and IA are remarkably different; 45% of males (39 subjects) and 26% of females (13 subjects) are classified as Type S, but 21% of males (18 subjects) and 34% of females (17 subjects) are classified as Type IA. The gender distributions by type are significantly different (Wilcoxon rank-sum test, $p = 0.0232$), and the differences in the gender distributions for Types S and IA are also significant (Wilcoxon rank-sum test, $p = 0.0241$). This result is in line with results from studies of gender differences in fairness considerations; namely, women usually prefer fairness, and men are more likely to favor efficiency (Selten & Ockenfels, 1998; Andreoni & Vesterlund, 2001; Dickinson & Tiefenthaler, 2002; Dufwenberg & Muren, 2006).

Second, there are significant differences in GSS Trust, social acceptance and compassion, as shown in Table 2.4. Subjects of Type S answer that most people can be trusted at the highest rate (40%), but Type IA subjects do so at the lowest (11%). Type IA

of you if they got a chance, or would they try to be fair?" (1: "would take advantage", 0: "would try to be fair"); and 3) GSS Help: "Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?" (1: "try to be helpful", 0: "just look out for themselves"). We used the Japanese translation of these questions.

[18]Cooperativeness refers to an individual's attitude about others, namely, whether he or she is agreeable to others or is a self-centered person. We choose this dimension only because the motivations for punishment, that is, reciprocity and fairness considerations, are social preferences.

Table 2.4: Results of the Questionnaire.

Variable	Type S	Type R	Type IA	Type O
(1) GSS (Percentage who are trusting)				
GSS Trust	40%	35%	*11%*	22%
GSS Fair	56%	50%	51%	33%
GSS Help	33%	28%	20%	22%
(2) Cooperativeness (Average score out of 5 points)				
Social acceptance	3.83	3.88	*4.26*	3.78
Empathy	2.23	2.18	2.34	2.00
Helpfulness	3.54	3.53	3.49	3.56
Compassion	3.46	3.28	*2.91*	3.78
Pure-hearted conscience	3.46	3.48	3.51	3.89

reports a significantly lower rate on GSS Trust than Types S and R (Wilcoxon rank-sum test, $p = 0.0045$ and $p = 0.0248$, respectively). Despite neither cooperating nor punishing, Type S exhibits the highest level of trust, which is similar to the finding in Gächter, Herrmann, and Thöni (2004) that trust is not significantly correlated with cooperation. Furthermore, regarding social acceptance and compassion scores, Type IA subjects show significant differences from other types. Their social acceptance scores are higher than those of Types S and R at the 10% level (t-test; compared with Type S: $p = 0.0743$, with Type R: $p = 0.0815$, and with Type O: $p = 0.1462$). Additionally, they have significantly lower compassion scores than Types S and O at the 10% level (t-test; compared with Type S: $p = 0.0654$, with Type R: $p = 0.2127$, and with Type O: $p = 0.0935$). One possible conjecture is that Type IA punishes not others' behaviors, namely, free-riding, but the outcome, namely, payoff inequality, because Type IA tends to have low trust in others (low GSS Trust) and accepts free-riding that does not lead to unequal outcomes (high social acceptance scores).

2.4 Summary and Discussion

Previous studies have addressed which of the two models—reciprocity or inequality aversion—offers a more plausible explanation for punishment motivations in cases where there is no future benefit from punishing others. However, in this chapter, we attempted to distinguish between these two motivations for punishment and to identify heterogeneity by introducing noise through an uncertainty factor and running individual-level regressions of punishment decisions. The results show that there are four types of subjects espousing different motivations for punishment: the self-interested type (Type S), the reciprocal type (Type R), the inequality-averse type (Type IA), and the "other type" (Type O). Individuals motivated to punish based on inequality aversion exist in similar proportion to those motivated by reciprocity.

We find several characteristics of each type, which supports the plausibility of our classification. First, in the case of Type R, subjects punish uncooperative behaviors because they are the most cooperative. Moreover, in the random income game, Type R is the most affected by intentionality among all types. Second, on the other hand, Type IA subjects change their punishment of others due to changes in intentionality less than Type R because the former's concerns about inequality in payoffs are stronger than others' behaviors. Last, subjects who do not punish at all in the baseline game, namely, Type S subjects, are the least cooperative. Additionally, we find that there are differences by

gender and social characteristics. Female subjects are more likely to be Type IA, and male subjects are more likely to be Type S. This result is in line with results from studies of gender differences in fairness considerations—namely, that women usually prefer fairness and men are more likely to favor efficiency. Type IA subjects are less generally trusting but have high social acceptance.

In the context of costly punishments, reciprocity, the motivation for punishing others' contribution decisions (or intentions), has been researched most frequently, and punishments based on people's reciprocity are verified to have a positive effect on enforcing cooperation. However, in our results, inequality aversion exists alongside reciprocity. Additionally, when reciprocity and inequality aversion can be distinguished through uncertainty, some people want to reduce inequality regardless of whether the target has good or bad intentions. Our result indicates the possibility of misunderstandings between punishments for outcomes and antisocial punishments, which can potentially hinder cooperation. This emphasizes the importance of ensuring clarity and precision in the implementation of punishments for outcomes, particularly in circumstances with uncertainty.

Nevertheless, the characteristics of each type in this chapter indicate another interpretation of Type R and Type IA behaviors. First, the existence of Type IA implies that the distribution of outcomes is as important as encouraging cooperation, especially when there is uncertainty, as in the real world. Second, the result that Type IA is least generally trusting means that reducing inequality in outcomes is more reasonable than cooperating with untrustworthy strangers. Third, regarding the results for the Type R subjects, individuals who punish based on reciprocity also consider payoff inequality. Thus, when there is uncertainty, as in the real world, it is insufficient to deal with costly punishments in terms of reciprocity, as has been done in most previous studies, because not only subjects of Type IA but also those of Type R care significantly about outcome inequality. That is, it is necessary to consider outcome inequality itself, even if it arises through a fair process or good intentions.

Of course, in addition to reciprocity and inequality aversion, other social preferences can affect the punishment decision. For example, Charness and Rabin (2002) and Engelmann and Strobel (2004) show that efficiency maximization is a stronger motivator than inequality aversion. In the case of our experiment, efficiency maximization could result in free riders being punished to increase the overall efficiency within the experimental "society" by creating a culture of cooperation. However, efficiency maximization is hard to examine in our study because of stranger matching and the uncertainty factor. Even if one's punishment induces cooperation by the target in the following rounds, one is unlikely to meet that target again in the same group. Moreover, even though overall cooperation increases across rounds, the uncertainty factor is likely to inhibit the benefits of cooperation. Competitiveness, envy, and spite could be possible motivations for punishment. Bault et al. (2008) suggest that people want to win in terms of their payoff ranking, and Casal et al. (2012) and Zizzo and Oswald (2001) show that envy can cause individuals to punish others. Hilbe and Traulsen (2012) show that some people punish spitefully regardless of others' behavior. In the cases of competitiveness and envy, the subject punishes a target who has earned a higher payoff than himself or herself to raise his or her payoff ranking. In our study, however, it is difficult to clearly distinguish these motivations from inequality aversion because inequality-averse subjects also punish others who have higher payoffs than themselves. Therefore, designing a new experimental setting, which would allow us to identify more diverse motivations such as those mentioned above, remains part of our future work.

In conclusion, our study implies that people who are willing to punish others without

future expected benefits but have different motivations illustrate the various purposes of punishment in society: to enhance cooperation and reduce inequality. Furthermore, our methodology, which classifies subjects by type and analyzes the characteristics of each type, namely, researching the heterogeneity in motivations for punishment, may be an effective tool for understanding the various sides of the punishment debate. This result suggests that researching heterogeneity in the population is useful for understanding society more deeply.

Chapter 3

Heterogeneity in Individual Fairness Ideals

3.1 Introduction

As mentioned in Chapter 1, Cappelen et al. (2007) and other literature demonstrate the pluralism of fairness ideals. In this chapter, we attempt to more accurately estimate the distribution of adherence to fairness ideals. Building upon the works of Cappelen et al. (2007) and Almås et al. (2010), we introduce modifications to the experiments and estimation models.

First, we employ the finite mixture model (Moffatt, 2015). Previous studies have estimated the distribution using mixed logit with repeated choice (Revelt & Train, 1998) because the data on the dictator's offer are discrete. In these experiments, each subject chooses their share from given options, where, for example, the choice set includes multiples of 50. In contrast, our experiments allow each subject to input their distribution of total production within the integer range. Given this significant expansion of subjects' options, we opt for the finite mixture model (Moffatt, 2015).

Second, we introduce a pure egoist into the choice model proposed by Cappelen et al. (2007) – individuals who prefer to take the entire product for themselves. There is a potential risk in exclusively considering fairness ideal types without accounting for this egoistic preference. In the context of the dictator game, the dominant strategy in distribution decisions is to distribute nothing to others. Therefore, we modify the model of Cappelen et al. (2007) by assuming four types: strict egalitarians, libertarians, liberal egalitarians, and pure egoists. This modification aims to acknowledge the possibility of risk hedges in our experiments, where decision-making may lean towards selfish choices due to the risk that one's decision may not yield favorable outcomes through the strategic method. Moreover, this perspective aligns with the notion of profit pursuit, as argued by Rodriguez-Lara and Moreno-Garrido (2012) and Ubeda (2014). Under these circumstances, individuals who choose not to adhere to a fairness ideal by strongly considering risk or their own profit could be characterized as egoists. However, those who opt for a distribution close to one of three fairness ideals, despite the possibilities of risk hedges, may be classified as one of the fairness ideal types.

Third, additionally, we propose another modified model that assumes the parameter representing the weight on the fairness ideal (β in Cappelen et al. (2007)) can reveal the characteristics of each fairness ideal. This contrasts with Cappelen et al. (2007)'s assumption that, since β is an individual characteristic, they can estimate only the distribution of β. Based on our assumption for β, we set different parameters (β^{SE} β^{L} and β^{LE}) in

each fairness ideal model. Through these parameters, we can compare how important each group of subjects following the same fairness ideal considers their ideal to be. Since we cannot conduct our analysis using the original data ofCappelen et al. (2007), we design and conduct our own experiments based on the experiment of Almås et al. (2010), which adds the real-effort task to Cappelen et al. (2007)'s experimental design.

The main findings of this study are as follows: First, around 40% of subjects are classified as egoists, representing a substantial majority. However, despite this majority, the three fairness ideal types proposed by Cappelen et al. (2007) still coexist. This can be taken as confirmation of the existence of people with fairness ideals, even when the purely egoistic type is included. This finding may pose a challenge to the argument presented by Rodriguez-Lara and Moreno-Garrido (2012) and Ubeda (2014) that people choose fairness ideals in a selfish manner; namely, individuals tend to change their fairness ideals to maximize their own payoffs rather than adhere to a particular ideal.

Second, the model that separates the weights (β) given to fairness ideals reveals the characteristics of each fairness ideal type. Strict egalitarians assign the highest weight to their ideal, followed by liberal egalitarians, with libertarians having the lowest weight among the three types. This result indicates that the importance of the fairness ideal varies depending on which ideal people adhere to, suggesting that when considering fair distribution, the more factors attributable to the individual—such as individual luck and effort—are included, the lower the weight placed on the ideal tends to be.

Third, we use posterior type probabilities to compare Cappelen et al. (2007)'s model with our modified model and assess the fitness of the estimates. We find that our own model, which includes egoists, predicts the observations more accurately and, thus, better accounts for the heterogeneity of fairness ideals in the population.

This chapter proceeds as follows. In Section 3.2, we describe the experimental design based on Cappelen et al. (2007) and Almås et al. (2010). Section 3.3 provides the details of the choice model and the three fairness ideals, and Section 3.4 explains the estimation models of Cappelen et al. (2007) and our modifications. Section 3.5 presents the results, and Section 3.6 gives a summary and conclusions.

3.2 The Experiments

Our experiments are based on the experiments of Almås et al. (2010) that add a real-effort task to Cappelen et al. (2007)'s experiments. The experiments consist of two phases: the production phase with a real-effort task and the distribution phase.

After seating all participants, we distributed the instructions for the production phase only, and the computer read the instructions aloud. In this phase, each participant participated in the real-effort task, which was the slider task of Gill and Prowse (2012). This slider task can be performed simply on a computer in the laboratory, and it is easy to understand and less affected by individual skills (Charness, Gneezy, & Henderson, 2018). At the beginning of this task, all 48 sliders were positioned at 0. By using the mouse, the subject could move each slider to any location between 0 and 100. The subject obtained points in the task based on the number of sliders positioned at 50 at the end of the allotted time. We first conducted an exercise task, the allotted time for which was 120 seconds. Next, the new slider task was conducted for 120 seconds in sessions 1–3 and for 150 seconds in sessions 4–7. When the time was up, the result screen was shown. It included the points from the slider task (we call these point individual i's effort, q_i); the rate of return, which was randomly assigned to be high or low (we call this i's productivity, $a_i \in \{2, 4\}$); and the individual's output (x_i), which was the product of individual productivity a_i and

Table 3.1: Summary of Sessions

Session number	Time limit of the production phase	Average efforts	Average time per correct slider	Total number of subjects
1 – 3	120 seconds	15.3	7.84	80
4 – 15	150 seconds	17.25	6.96	282
Total				362

effort q_i ($x_i = a_i q_i$). To prevent productivity from affecting the performance on the slider task, we presented each subject's productivity on the result screen after the tasks were finished.

After all participants checked their results from the production phase, the instruction for the distribution phase were distributed and read by the computer in the same way. In this phase, subjects with different productivity levels were randomly paired and made distribution decisions. The sum of own output (x_i) and the partner's output (x_j) became the group's total product ($X_i = x_i + x_j$), and each subject decided how to distribute the total group product to himself or herself and the partner. In this phase, each subject was informed of his or her own and the partner's points from the task (q_i, q_j), the rate of return (a_i, a_j), and output (x_i, x_j). This distribution decision was repeated for six rounds with perfect stranger matching, and the partner's decision was not provided. All participants knew that both subjects would make decisions independently as a dictator and that they would not be matched with the same partner again. The payment was the sum of two distribution results selected according to the following procedure: 1) we randomly selected two rounds out of the six, 2) we randomly selected one of the two subjects' distribution decisions in the two selected rounds, 3) we summed the results of the two selected distribution decisions and paid each subject.

The experiments were conducted at Waseda University from October 2019 to May 2023. We ran fifteen sessions with 18 to 30 subjects in each session, and all 362 participants were students in various majors from Waseda University (Table 3.1)[1]. The experiments were computerized with z-Tree (Fischbacher, 2007). Each session lasted 60 minutes, and the show-up payment was ¥1000 (≈ \$7.9), and average earnings were ¥1345 (≈ \$10.6)[2].

3.3 The Choice Model

In the choice model of Cappelen et al. (2007), individual i trades off the amount for himself or herself (y_i) and the amount corresponding to his or her ideal fair distribution ($m^{k(i)}$), as in the following utility function.

$$V_i^{k(i)}(y_i) = y_i - \beta_i \frac{(y_i - m^{k(i)})^2}{2X_i} \tag{3.1}$$

where y_i is the offer for himself or herself (and $X_i - y_i$ is for his or her partner); $m^{k(i)}$ is a fairness ideal, which is the amount that i considers his or her fair income; $\beta_i (\geq 0)$ is the

[1] In Sessions 1-3, certain participants earned comparatively fewer points on the slider task due to the imposed time limit. To mitigate the potential influence of time constraints on decision-making, we extended the time limit to 150 seconds from Session 4.

[2] We convert Japanese yen into US dollars at the exchange rate of \$1=¥127, which was the average rate at the time that the experiments were conducted.

weight assigned to fairness; and X_i is the total product of i's group to be distributed. i's optimal offer for himself or herself is as follows by the first-order condition:

$$y_i^* = m^{k(i)} + \frac{X_i}{\beta_i}. \tag{3.2}$$

If $\beta_i = 0$, an individual expresses no interest in fairness and chooses to take all of X_i. However, the upper bound of y_i is X_i ($y_i \leq X_i$), and there exists $\beta_i \neq 0$ that derives the optimal $y_i^* = X_i$, the selfish choice. In this case, the optimal choices of a strict egalitarian and an egoist become indistinguishable in the results of our experiments. Consequently, there is a potential risk that the estimation results from Cappelen et al. (2007) might overlook the existence of the egoist in the estimation.

There are three fairness ideals as mentioned in Section 3.1. First, by the strict egalitarian (SE) ideal, the outcome must be distributed equally ($m^{SE(i)}$). Second, according to the libertarian (L) ideal, each individual has the right to his or her own output because both his or her productivity and effort are due to the individual ($m^{L(i)}$). By the liberal egalitarian (LE) ideal, each individual is responsible only for his or her choice or effort, and the effects of luck should be excluded ($m^{LE(i)}$). In addition, we include the egoistic (EGO) ideal in our new model, according to which taking all is a reasonable choice when possible ($m^{EGO(i)}$). The fair shares under each fairness ideal are as follows:

$$\begin{aligned} m^{SE(i)} &= \frac{X_i}{2} \\ m^{L(i)} &= a_i q_i \\ m^{LE(i)} &= \frac{q_i}{q_i + q_j} X_i \\ m^{EGO(i)} &= X_i \end{aligned}$$

where q_i is the level of effort (in our experiment, total points from the slider task), a_i is productivity (in our experiment, the rate of return, high or low), and X_i is the total product of group i.

3.4 Finite Mixture Model Estimation

In our experiments, since the distribution decisions of subjects are continuous variables, mixed logit with repeated choices (Revelt & Train, 1998), which is used in Cappelen et al. (2007) and Almås et al. (2010), cannot be applied for our estimation. Thus, the estimation is conducted by the finite mixture model (Moffatt, 2015) [3]. This model estimates the distribution of multiple behavioral models by adapting them to the estimation. There are various approaches to the estimation of the finite mixture model, but Moffatt (2015)'s approach is adopted in our study as follows: First, the models of the three fairness ideals are decided, and a label is assigned to each. Second, a parametric model is specified for the behavior corresponding to each ideal. Third, the parameters of these three models are estimated jointly, along with the "mixing proportion"—the proportion of the subjects who follow each ideal. Finally, we determine the posterior probability of each subject following each ideal[4]. Specifically, we modify one assumption from Cappelen et al. (2007).

[3] To verify whether the distribution obtained represented the Global Maximum, we examined the Maximum Likelihood Estimates by randomly varying the starting values during our estimation process and selected the result with the highest maximum likelihood estimate.

[4] However, we cannot identify any subject's fairness ideal with certainty.

In Cappelen et al. (2007), they assume that the parameter (β_i) is an individual variable, so they estimate only the approximate distribution of β_i by a log-normal distribution. However, we assume that the importance that a person assigns to fairness can be considered a characteristic of the group of subjects who follow the same fairness ideal. Thus, we separate the parameter β by fairness ideal, such that β^{SE}, β^{L}, and β^{LE}. In this section, we describe the estimation models.

3.4.1 The Previous Empirical Model (Model P) of Cappelen et al. (2007)

First, we estimate the mixing proportion of the three fairness ideals by adopting Cappelen et al. (2007)'s method that estimates the integrated β. Cappelen et al. (2007) assume that β_i follows a log-normal distribution, and the average value of $\log(\beta)$ is estimated. However, since our study assumes the existence of $\beta_i > 0$ in which the egoistic offer $y_i^* = X_i$ becomes optimal as explained in Section 3.3, we estimate β without applying the log. Therefore, from the choice model in Section 3.3, the optimal offer y_i of each type including the error term is as follows:

1) Type 1 (Strict egalitarian, SE)

$$y_i = \frac{X_i}{2} + \frac{X_i}{\beta} + \epsilon_{1,i}^P. \tag{3.3}$$

2) Type 2 (Libertarian, L)

$$y_i = a_i q_i + \frac{X_i}{\beta} + \epsilon_{2,i}^P. \tag{3.4}$$

3) Type 3 (Liberal egalitarian, LE)

$$y_i = \frac{q_i}{q_i + q_j} X_i + \frac{X_i}{\beta} + \epsilon_{3,i}^P. \tag{3.5}$$

We assume that all three errors have the same variance as follows: $V(\epsilon_{1,i}^P) = V(\epsilon_{2,i}^P) = V(\epsilon_{3,i}^P) = \sigma^2$.

The mixing proportions of each ideal are p_1^P, p_2^P, and p_3^P, and the likelihood contribution of each subject i is as follows:

$$\begin{aligned} L_i^P = p_1^P \frac{1}{\sigma}\phi\left(\frac{y_i - \frac{X_i}{2} - \frac{X_i}{\beta}}{\sigma}\right) + p_2^P \frac{1}{\sigma}\phi\left(\frac{y_i - a_i q_i - \frac{X_i}{\beta}}{\sigma}\right) \\ + (1 - p_1^P - p_2^P)\frac{1}{\sigma}\phi\left(\frac{y_i - \frac{q_i}{q_i + q_j} X_i - \frac{X_i}{\beta}}{\sigma}\right). \end{aligned}$$

where ϕ is the probability density function of the normal distribution. The parameters to be estimated are β, σ, p_1^P, and p_2^P, and we conduct log-likelihood estimation.

3.4.2 Our Second Modified Model with Egoism (Model M1)

As mentioned in the introduction, the distribution decisions of subjects are basically based on the dictator game. Thus, it is questionable to not postulate the egoistic behavior of always takes all of the total product because it is a reasonable decision in the dictator game. In this section, we suggest an modified model that adds an egoist. We adopt the assumption of β in Cappelen et al. (2007) which estimates the average value of β because each individual who has the fairness ideal would place a different weight on his or her ideal. Therefore, we set the new parameter $\bar{\beta}$ for the subjects who have fairness ideals. From our second modified choice model, the optimal offer y_i of each type including the error term is as follows:

1) Type 1 (Strict egalitarian, SE)

$$y_i = \frac{X_i}{2} + \frac{X_i}{\bar{\beta}} + \epsilon_{1,i}^{M1}. \tag{3.6}$$

2) Type 2 (Libertarian, L)

$$y_i = a_i q_i + \frac{X_i}{\bar{\beta}} + \epsilon_{2,i}^{M1}. \tag{3.7}$$

3) Type 3 (Liberal egalitarian, LE)

$$y_i = \frac{q_i}{q_i + q_j} X_i + \frac{X_i}{\bar{\beta}} + \epsilon_{3,i}^{M1}. \tag{3.8}$$

4) Type 4 (Egoist, EGO)

$$y_i = X_i + \epsilon_{4,i}^{M1}. \tag{3.9}$$

where $\bar{\beta}$ is a parameter that represents subjects' weight assigned to their fairness ideal: liberal egalitarianism or libertarianism. The egoist does not have a fairness ideal, and takes all of the total product X_i, that is, $\bar{\beta}_i = 0$. We assume that all three errors have the same variance as follows: $V(\epsilon_{1,i}^{M1}) = V(\epsilon_{2,i}^{M1}) = V(\epsilon_{3,i}^{M1}) = V(\epsilon_{4,i}^{M1}) = \sigma^2$.

The mixing proportions of each ideal are p_1^{M1}, p_2^{M1}, p_3^{M1}, and p_4^{M1}, and the likelihood contribution of each subject i is as follows:

$$\begin{aligned} L_i^{M1} =& p_1^{M1} \frac{1}{\sigma} \phi \left(\frac{y_i - \frac{X_i}{2} - \frac{X_i}{\bar{\beta}}}{\sigma} \right) + p_2^{M1} \frac{1}{\sigma} \phi \left(\frac{y_i - a_i q_i - \frac{X_i}{\bar{\beta}}}{\sigma} \right) \\ &+ p_3^{M1} \frac{1}{\sigma} \phi \left(\frac{y_i - \frac{q_i}{q_i + q_j} X_i - \frac{X_i}{\bar{\beta}}}{\sigma} \right) + (1 - p_1^{M1} - p_2^{M1} - p_3^{M1}) \frac{1}{\sigma} \phi \left(\frac{y_i - X_i}{\sigma} \right). \end{aligned}$$

where ϕ is the probability density function of the normal distribution. The parameters to be estimated are $\bar{\beta}, \sigma, p_1^{M1}$, p_2^{M1}, and p_3^{M1}, and we conduct log-likelihood estimation.

3.4.3 Our Modified Model with Separate Parameters by Fairness Ideal (Model M2)

Next, in addition to our modified model 1, we claim Cappelen et al. (2007)'s assumption of the parameter β as explained above; namely, in our context, the importance that a person assigns to fairness can be considered a characteristic of each fairness ideal. Thus, people who follow the same fairness ideal have their own parameters, such that β^{SE} for the strict egalitarian, β^{L} for the libertarian, and β^{LE} for the liberal egalitarian. Therefore, from our modified choice model, the optimal offer y_i of each type including the error term is as follows:

1) Type 1 (Strict egalitarian, SE)

$$y_i = \frac{X_i}{2} + \frac{X_i}{\beta^{SE}} + \epsilon_{1,i}^{M2}. \tag{3.10}$$

2) Type 2 (Libertarian, L)

$$y_i = a_i q_i + \frac{X_i}{\beta^{L}} + \epsilon_{2,i}^{M2}. \tag{3.11}$$

3) Type 3 (Liberal egalitarian, LE)

$$y_i = \frac{q_i}{q_i + q_j} X_i + \frac{X_i}{\beta^{LE}} + \epsilon_{3,i}^{M2}. \tag{3.12}$$

4) Type 4 (Egoist, EGO)

$$y_i = X_i + \epsilon_{4,i}^{M2}. \tag{3.13}$$

We assume that all three errors have the same variance as follows: $V(\epsilon_{1,i}^{M2}) = V(\epsilon_{2,i}^{M2}) = V(\epsilon_{3,i}^{M2}) = V(\epsilon_{4,i}^{M2}) = \sigma^2$.

The mixing proportions of each ideal are p_1^{M2}, p_2^{M2}, p_3^{M2}, and p_4^{M2}, and the likelihood contribution of each subject i is as follows:

$$\begin{aligned} L_i^{M1} =& p_1^{M2} \frac{1}{\sigma} \phi \left(\frac{y_i - \frac{X_i}{2} - \frac{X_i}{\beta^{SE}}}{\sigma} \right) + p_2^{M2} \frac{1}{\sigma} \phi \left(\frac{y_i - a_i q_i - \frac{X_i}{\beta^{L}}}{\sigma} \right) \\ &+ p_3^{M2} \frac{1}{\sigma} \phi \left(\frac{y_i - \frac{q_i}{q_i + q_j} X_i - \frac{X_i}{\beta^{LE}}}{\sigma} \right) + (1 - p_1^{M2} - p_2^{M2} - p_3^{M2}) \frac{1}{\sigma} \phi \left(\frac{y_i - X_i}{\sigma} \right). \end{aligned}$$

where ϕ is the probability density function of the normal distribution. The parameters to be estimated are $\beta^{SE}, \beta^{L}, \beta^{LE}, \sigma, p_1^{M2}$, p_2^{M2}, and p_3^{M2}, and we conduct log-likelihood estimation.

3.4.4 Posterior Type Probabilities

According to Moffatt (2015) and Cappelen et al. (2011), the posterior probability that each subject follows each fairness ideal can be calculated using the estimated parameters from estimations. Among the two type of calculations, we use the calculation equation from Moffatt (2015) in this paper. For example, the equation for calculating the posterior probability of following each fairness ideal using the estimated value from Model P is as follows.

1) Strict egalitarian (SE)

$$P(i = \text{SE}|y_{i1}, ..., y_{iT}) = \frac{p_1^P \prod_{t=1}^{T} \frac{1}{\sigma}\phi\left(\frac{y_i - \frac{X_i}{2} - \frac{X_i}{\beta}}{\sigma}\right)}{Li}$$

2) Libertarian (L)

$$P(i = \text{L}|y_{i1}, ..., y_{iT}) = \frac{p_2^P \prod_{t=1}^{T} \frac{1}{\sigma}\phi\left(\frac{y_i - a_i q_i - \frac{X_i}{\beta}}{\sigma}\right)}{Li}$$

3) Liberal egalitarian (LE)

$$P(i = \text{LE}|y_{i1}, ..., y_{iT}) = \frac{(1 - p_1^P - p_2^P) \prod_{t=1}^{T} \frac{1}{\sigma}\phi\left(\frac{y_i - \frac{q_i}{q_i+q_j}X_i - \frac{X_i}{\beta}}{\sigma}\right)}{Li}$$

This calculation is based on Bayesian rules. In the case of Model M1, we use the variables $\bar{\beta}, \sigma$, p_1^{M1}, p_2^{M1}, and p_3^{M1}, and in the case of Model M2, we include the variables β^{SE}, $\beta^{L}, \beta^{LE}, \sigma, p_1^{M2}$, p_2^{M2}, and p_3^{M2} in the equation for the posterior probability. Using each subject's six decisions in the distribution phase, we calculate the posterior probability of following each fairness ideal and the fairness ideal with the highest probability value is the type of each subject. When all of the posterior types of each subject are derived, the predictions of each subject's optimal offer (y_i) can be derived from the type using the estimated parameters. We compare the predictions of optimal offers from the posterior type and the observed offers and examine which model's estimation better fits the observations.

3.5 Results

3.5.1 Estimation

Before proceeding with the estimations described in Section 3.4, we review the observations of individuals' decisions. In Figure 3.1[5], the left panel depicts the relationship between individual decisions on the vertical axis and the individual's proportion of production, i.e., the libertarian ideal share, on the horizontal axis. On the right panel, the figure illustrates the relationship between individual decisions on the vertical axis and the individual's proportion of effort, i.e., the liberal egalitarian ideal share, on the horizontal axis. As

[5]The format of Figure 3.1 is roughly based on Figure 3 in Frohlich, Oppenheimer, and Kurki (2004).

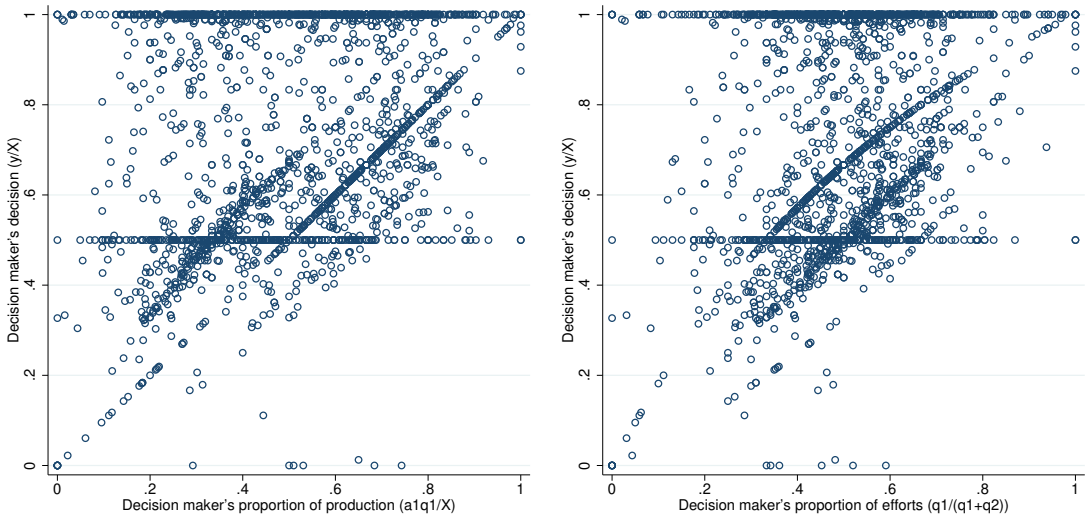


Figure 3.1: Scatter Graphs of Distribution Decision

shown, a distinct cluster consistently claims the entire production for themselves ($y/x = 1$), regardless of an individual's luck and effort, which contribute to the size of total production, while another group consistently opts for an equal division ($y/x = 0.5$). Figure 3.1 also displays the libertarian and liberal egalitarian distributions on a 45-degree line. This observation suggests that our modified model incorporating egoism could potentially provide a more accurate fit, aligning with the focus of our study, as detailed in Section 3.4.

Now by using the estimation models described in Section 3.4, we estimate the weight assigned to fairness (β, $\bar{\beta}$, β^{SE}, β^{L}, and β^{LE}) and the proportion of each type in the population as shown in Table 3.2.

Firstly, from the estimation results of Model P, the proportion of strict egalitarians is 0.3739, which is within the range of estimates from previous studies: 0.18 in Cappelen, Sørensen, and Tungodden (2010), 0.301 in Cappelen et al. (2011), 0.365 in Almås et al. (2010), and 0.435 in Cappelen et al. (2007). Since the proportions of libertarians and liberal egalitarians vary across socioeconomic conditions in previous studies, it can be said that the results of this study are also similar to those of previous studies.

Secondly, the fairness ideal distribution of Models M1 and M2 reveals a distinct contrast with the outcomes observed in Model P. Models M1 and M2, modified to incorporate egoism, result in a reduction of the proportion of strict egalitarians from 0.3739 in Model P to 0.19 and 0.2147, and a decrease in the proportion of libertarians from 0.3974 in Model P to 0.1906 and 0.1831, respectively. Notably, the majority type in Models M1 and M2 is the egoist, comprising 0.3683 and 0.3695 of the population, respectively. Consequently, the existence of egoists, who take the entire product, cannot be understated. The significant prevalence of egoists, constituting 0.3683 in Model M1 and 0.3695 in Model M2, aligns with the arguments posited by Rodriguez-Lara and Moreno-Garrido (2012) and Ubeda (2014). These findings suggest that individuals may strategically adopt different ideals based on situational factors, aiming to maximize their payoffs rather than strictly adhering to a predefined fairness ideal. Given that more than one-third of subjects in our study exhibit egoistic behavior, these results may provide empirical support for the assertions made by Rodriguez-Lara and Moreno-Garrido (2012) and Ubeda (2014) regarding the prevalence

Table 3.2: MLEs from Three Models

	Model P	Model M1	Model M2
β	4.0429		
	(0.0364)		
$\bar{\beta}$		9.2533	
		(0.2958)	
β^{SE}			13.1353
			(0.8014)
β^{L}			7.8157
			(0.2474)
β^{LE}			9.2069
			(0.3413)
Proportion of SE	0.3739	0.1900	0.2147
	(0.0170)	(0.0139)	(0.0135)
Proportion of L	0.3974	0.1906	0.1830
	(0.0142)	(0.0130)	(0.0123)
Proportion of LE	0.2287	0.2511	0.2328
	(0.0161)	(0.0140)	(0.0136)
Proportion of EGO		0.3683	0.3695
		(0.0109)	(0.0109)
σ	18.2954	11.3246	11.2501
	(0.1626)	(0.0993)	(0.0980)
Log-Likelihood	-33894.788	-31166.482	-31131.127

Note: Standard errors are given in parentheses. SE: Strict Egalitarian, L: Libertarian, LE: Liberal Egalitarian, EGO: Egoist.

of such behaviors. In Section 3.5.3, we compare the fitness of these three models with the aim of identifying the model that best aligns with the observed data.

Thirdly, in Model M2, weights assigned to each fairness ideal, β^{SE}, β^{L} and β^{LE}, are considerably higher than β in Model P. In particular, strict egalitarians are estimated to have the highest β ($\beta^{SE} = 13.1353$), followed by liberal egalitarians ($\beta^{LE} = 9.2069$), and libertarians have the lowest β ($\beta^{L} = 7.8157$) among these three types. This result may indicate that the more strict one's fairness ideal is — that is, the less one incorporates factors attributed to the individual, such as luck and effort, into the distribution decision — the less one tends to claim a self-interest premium (X_i/β_i in equation 3.2) to favor oneself. By separating the beta by types, we can discern the difference in the amount of weight attached to each fairness ideal as a characteristic of each type.

These estimation results show that the rational choice in the dictator game, to take all of the total product, should not be overlooked. Nevertheless, in our modified models, which incorporate both egoism and three fairness ideals, we can confirm that more than half of the population aligns with at least one fairness ideal, and the willingness to distribute according to those ideals is not vulnerable.

Table 3.3: Examples of Posterior Type Probability

	Model P			Model M1				Model M2			
ID	SE	L	LE	SE	L	LE	EGO	SE	L	LE	EGO
1	0.016	0.984	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
166	0.167	0.416	0.417	0.000	0.948	0.000	0.052	0.000	0.964	0.002	0.036
222	0.330	0.390	0.280	0.908	0.091	0.001	0.000	0.903	0.096	0.001	0.000
358	0.002	0.984	0.014	0.023	0.505	0.472	0.000	0.097	0.601	0.302	0.000

Note: SE: Strict Egalitarian, EGO: Egoist, L: Libertarian, LE: Liberal Egalitarian

3.5.2 Distribution of the Fairness Ideals by Posterior Type Calculation

Next, to compare the fitness of each model, it is necessary to predict the type of each subject based on the estimated parameters. Using the calculation of posterior type probability presented in Section 3.4.4, we can obtain the probability that each subject is classified into each fairness ideal type. Table 3.3 shows some examples in which the subject type is determined in this way.

For example, in the case of subject ID 1, when the posterior type is calculated with the estimated parameters of Model P, the probability of espousing the libertarian ideal is the highest at 98.4 percent, but when we use the parameters of Models M1 and M2, the probability of being an egoist is close to 100 percent. In the case of subject ID 166, in Model P, the probabilities of espousing the libertarian ideal and the liberal egalitarian ideal are 41.6 percent and 51.7 percent, respectively, but in Models M1 and M2, the probabilities of espousing the libertarian ideal are the highest, at 94.8 percent and 96.4 percent, respectively. In the case of subject ID 222, the posterior probabilities of the three types are similar: 30.0 percent for the strict egalitarian ideal, 39.0 percent for the libertarian ideal, and 28.0 percent for the liberal egalitarian ideal, such that he or she is rarely classified as a specific type. However, in Models M1 and M2, this subject is a strict egalitarian with a probability of 90.8 percent and 90.3 percent, respectively. Finally, in the example of subject ID 358, the highest probability is shown for the libertarian ideal in all three models. However, the probability of espousing the liberal egalitarian ideal in Model P is different from the probabilities for other models, namely, only in Model P is there no probability of espousing ideals other than the libertarian ideal.

Some subjects do not show a markedly high probability of holding one ideal, such as subject ID 166, who has a similar probability of being classified as a strict egalitarian and a liberal egalitarian in Model P in Table 3.3. However, since we calculate the posterior type from data on six decisions, most of the subjects have the highest posterior probability for one ideal. Therefore, each subject is classified as the type with the highest probability, and its distribution is presented in Table 3.4.

As we see from the results in Table 3.4, there is not much difference between the distribution derived by posterior type calculation and the distribution estimated by log-likelihood estimation (in Table 3.2). In other words, the types of subjects derived by posterior type calculation have validity.

Table 3.4: Comparison of the Distributions from Posterior Type Probability Calculation and from Estimation

Type	Model P		Model M1		Model M2	
	Number of subjects	Estimated proportion	Number of subjects	Estimated proportion	Number of subjects	Estimated proportion
SE	123 (34.0%)	0.3739	58 (16.0%)	0.1900	62 (17.1%)	0.2147
L	159 (43.9%)	0.3974	69 (19.1%)	0.1906	74 (20.4%)	0.1830
LE	80 (22.1%)	0.2287	92 (25.4%)	0.2511	82 (22.7%)	0.2328
EGO			143 (39.5%)	0.3683	144 (39.8%)	0.3695
Total	362		362		362	

3.5.3 Fitness Test

By using each subject's posterior type from the three estimation models and the parameters β, $\bar{\beta}$, β^{SE}, β^{L}, and β^{LE} of the three models, we can predict each subject's optimal offer y_i in the six distribution decisions under different a_j, q_j, and X_i. In this section, the distribution of predictions derived by the three models and the distribution of observations are compared, and we identify which model's prediction is most similar to the observations.

To compare the distribution of the three models' predictions, we conduct a t-test and a variance ratio test. We find no significant difference between the three models and observations in the t-test. However, in the variance ratio test, we cannot reject that the variance ratio of the observations and predictions of Model P is equal to 1 ($p = 0.0000$). Figure 3.2 shows the histograms and the normal distribution estimated through the mean and variance of the histogram. Figure 3.3 describes the estimated plots using the kernel density estimator[6]. Although all models appear to have similar means, Model P shows a different variance from the observations in Figure 3.3.

In addition, we examine subjects' questionnaire answers to the question "What principle did you use to divide the total product points in the distribution phase? Please write briefly". The results are as follows: 143 people who are classified as type EGO in our Model M1 wrote that they distributed to maximize their own profit, for example, "I distributed it to my advantage anyway", "I distributed thinking only about myself", "If the other does not know my decision, I want to take it all", and similar responses. In other words, these answers confirm that our estimations that include type EGO fit for the subjects' distribution decisions.

In conclusion, although the prediction of Model P does not differ substantially from the observations, the predictions of Models M1 and M2 have a better fit than Model P. Therefore, the estimations using Models M1 and M2 yield the result that there is a high proportion of egoists in population, and these models are more suitable for explaining the pluralism of fairness ideals in the population.

[6]Kernel density estimators approximate the density $f(x)$ from observations on x. Unlike a histogram, a kernel density estimator assigns a weight between 0 and 1—based on the distance from the center of the interval—and sums the weighted values. The function that determines these weights is called the kernel (Stata Corporation, 2005). We use the Epanechnikov kernel function to estimate the plots of the observations and the predictions of the three models, Model P, Model M1, and Model M2.

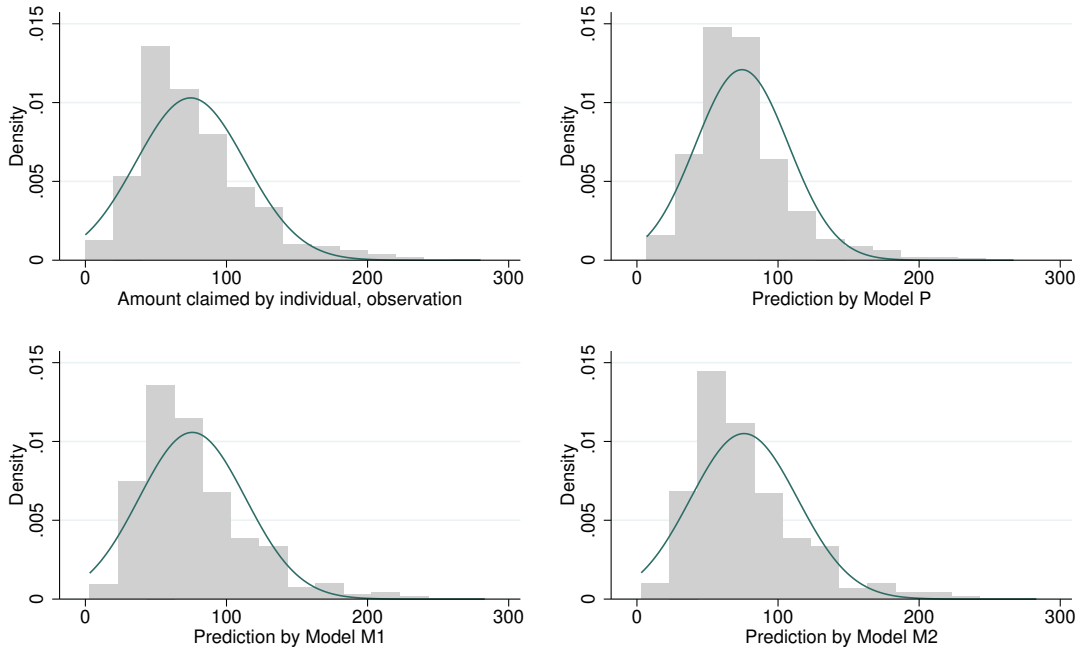


Figure 3.2: Histograms of Observations and Predictions of Three Models

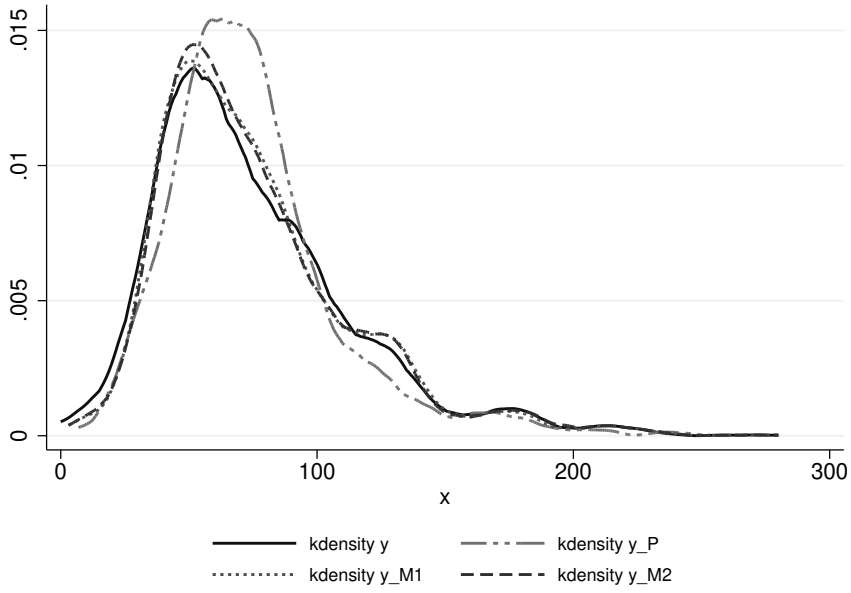


Figure 3.3: Kernel Density Estimations of Observations and Predictions of Three Models

3.5.4 Characteristics of Subjects with Different Fairness Ideals

What conditions in the experiments relate to people's fairness ideals? Are there differences in characteristics across subjects with different fairness ideals? In this section, we investigate the characteristics of subjects classified by means of the posterior types calculated using the parameters of our new Model M1. The reason for using Model M1 is that as

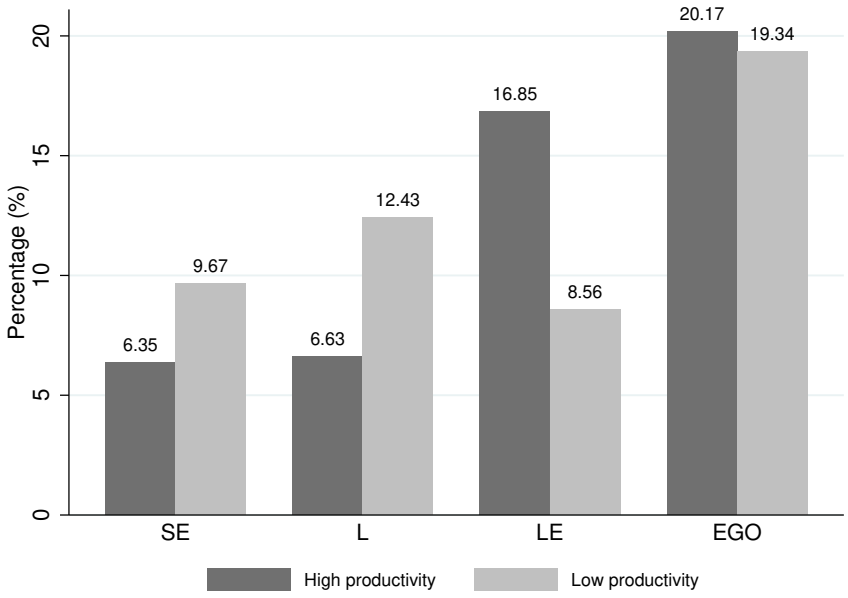


Figure 3.4: The Productivity Distributions by Fairness Ideals

described in Section 3.4.3, the assumption of Cappelen et al. (2007) on β, which estimates the average value of β because each individual who has a given fairness ideal would place a different weight on his or her ideal, is more simple and offers greater interpretability, and Model 1 is based on this assumption.

First, do the factors in the experiment, namely, randomly assigned productivity and effort exerted in the task, affect an individual's fairness type? According to the choice model and fairness ideals described in Section 3.3, libertarians consider their own output, which includes randomly assigned productivity, to be their own fair share, while liberal egalitarians believe that only individual efforts should be reflected in the distribution. Therefore, the following two conjectures could be possible: (1) subjects who are assigned high productivity may be libertarians because the allocation corresponding to the libertarian ideal reflects productivity and may yield the dictator a higher share that the allocations corresponding to the other ideals, and (2) subjects who show higher performance in the real-effort task in the production phase may follow the liberal egalitarian ideal because this ideal reflects only individual effort.

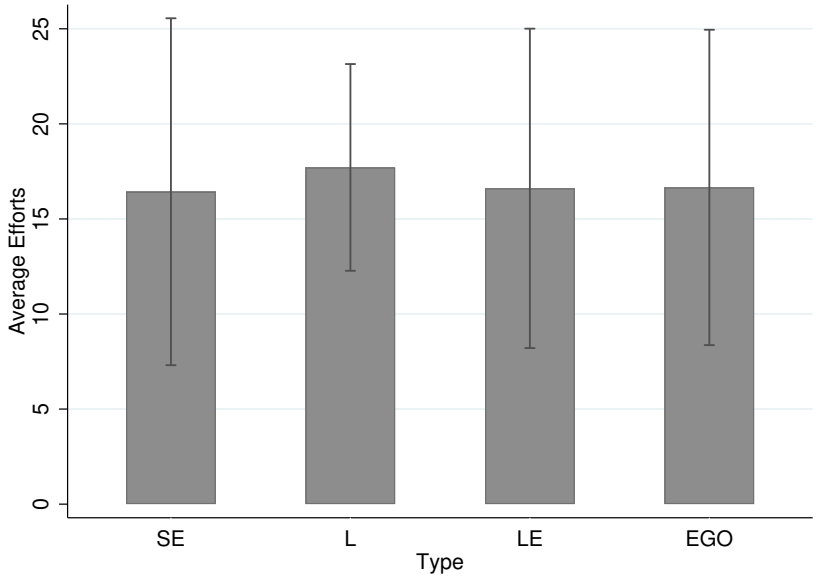
Table 3.5: Distribution of Productivity

Model M1	Number of subjects with high productivity	Number of subjects with low productivity	Total
SE	23	35	58
L	24	45	69
LE	61	31	92
EGO	73	70	143
Total	181	181	362

For conjecture (1), the type distribution among the people assigned each productivity

Table 3.6: Summary Statistics of Effort

Model M1	SE	L	LE	EGO
Average	16.43	17.71	16.61	16.66
Standard deviation	9.12	5.44	8.40	8.29
Minimum	1	7	0	0
Maximum	48	33	48	46



Note: Bars depict standard deviations.

Figure 3.5: The Average Efforts by Fairness Ideals

level is shown in Table 3.5 and Figure 3.4. Contrary to our conjecture, subjects who are assigned high productivity are classified as liberal egalitarians (LEs). Figure 3.4 illustrates the distribution of productivity based on fairness ideal types. Initially, the figure reveals that approximately two-thirds of libertarians exhibit low productivity, contrasting with high-productivity subjects who are more frequently categorized within the liberal egalitarian ideal. The difference in the productivity distributions between libertarians and liberal egalitarians is significant (rank-sum test, $p = 0.0000$). This result implies that high productivity, which is assigned randomly, may be related to the liberal egalitarian ideal, which insists that the distribution should not reflect random factors (in this case productivity) but rather only individual effort.

For conjecture (2), we compare the average effort by type. Table 3.6 shows summary statistics of effort for each fairness ideal type, and Figure 3.5 shows the average effort by type. We find, contrary to conjecture (2), that liberal egalitarians do not show higher performance than other types on average. On the other hand, as shown in Table 3.6 and Figure 3.5, libertarians show a significantly higher level of effort than other types (t-test, compared with the strict egalitarian: $p = 0.0085$, compared with the liberal egalitarian: $p = 0.0101$, and compared with the egoist: $p = 0.0094$). This result implies that subjects

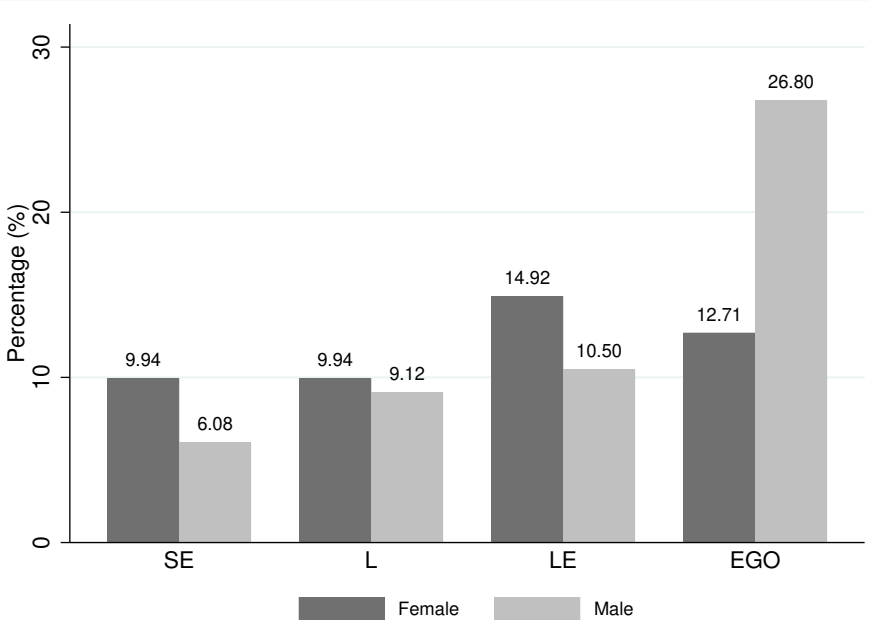


Figure 3.6: Gender Differences

who show higher performance in the real-effort task tend to be libertarians, who believe that the fair distribution should include not only individual effort but also the effect of random factors (here productivity), rather than liberal egalitarians, who believe that the fair distribution should be based only on individual effort.

In sum, these findings contradicting conjectures (1) and (2) may run contrary to the arguments of Rodriguez-Lara and Moreno-Garrido (2012) and Ubeda (2014) that people adopt different ideals depending on their situation in self-serving ways to maximize their payoffs. In other words, along with the fact that egoists exist within the population at a share of approximately 40 percent (as shown in Section 3.5.1), contrary to conjectures (1) and (2), the random factor (productivity) and the level of effort affect subjects' fairness ideals; that is, high productivity induces subjects to choose the libertarian ideal (L), and higher performance in the task induces subjects to follow the liberal egalitarian ideal (LE), implying that fairness ideals such as the libertarian and liberal egalitarian ideals could be considered a kind of distributional preference, not a result of self-serving bias.

The next characteristic related to fairness ideals is gender. The distribution of fairness ideals varies significantly by gender, as illustrated in Figure 3.6 (rank-sum test, $p = 0.0000$). Specifically, half of the male subjects are classified as egoists, while the four types are almost evenly distributed among females.

Table 3.7 is the result of multinomial logistic regression analysis that indicates the difference in the effect of each factor among subjects classified by fairness ideal type. The base outcome of the comparison is the strict egalitarian.

In Column 1 of Table 3.7, liberal egalitarians and egoists exhibit significant positive coefficients on productivity—0.5482 and 0.2306, respectively. Notably, the coefficient for libertarians is not statistically significant, indicating that subjects given low productivity tend to align with strict egalitarianism or libertarianism. On the other side, only libertarians display a positive coefficient on the level of effort in tasks—0.0203 at the 5% level. These findings imply the following: First, libertarians demonstrate the highest level of effort compared to other types. Second, when high productivity is assigned, subjects

Table 3.7: Multinomial Logistic Regression

Type	Variable	1	2
SE	Base outcome		
L	Productivity	−0.1082	−0.1031
		(0.0754)	(0.0756)
	Effort	0.0203*	0.0181
		(0.0092)	(0.0093)
	Gender		0.3660*
			(0.1491)
	Constant	0.1250	−0.0082
		(0.2663)	(0.2707)
LE	Productivity	0.5482***	0.5514***
		(0.0709)	(0.0710)
	Effort	0.0021	−0.0001
		(0.0089)	(0.0089)
	Gender		0.1791
			(0.1431)
	Constant	−1.2513***	−1.2973***
		(0.2671)	(0.2704)
EGO	Productivity	0.2306***	0.2523***
		(0.0646)	(0.0661)
	Effort	0.0033	−0.0038
		(0.0082)	(0.0084)
	Gender		1.2611***
			(0.1337)
	Constant	0.1770	−0.4379
		(0.2366)	(0.2489)

Note: Gender is a dummy variable, 0 for females, 1 for males. Standard errors are given in parentheses.
* $p < 0.5$; ** $p < 0.01$; *** $p < 0.001$.

are more inclined to align closely with the liberal egalitarian ideal than with the egoistic ideal, as suggested by the significantly higher coefficient for liberal egalitarians (0.5482) compared to that of egoists (0.2306).

Column 2 of Table 3.7 presents the results of the multinomial logistic regression, including the gender dummy variable (0 for female, 1 for male). The results of this regression closely resemble those in Column 1, except for libertarians. When using the strict egalitarian type as the reference group, high productivity shows a significant correlation with liberal egalitarian and egoistic ideals, supported by positive coefficients of 0.5514 and 0.2523, respectively. However, the level of effort is not significant across all types in this regression model.

In addition, male subjects exhibit a preference for the egoistic ideal, as indicated by

the highest coefficient for egoists at 1.2611 at the 0.1% significance level. This implies that male subjects are more likely to be egoistic types than female subjects. Furthermore, male subjects are also more likely to be libertarian types, although the magnitude is smaller and less significant, with a coefficient of 0.3660 at the 5% level. In other words, female subjects favor strict egalitarian and liberal egalitarian ideals over males. This aligns with the findings of Miller and Ubeda (2012) and Sharma (2015), who report that women tend to distribute based on effort. Additionally, our result indicating a preference for the strict egalitarian and the liberal egalitarian ideal among female subjects aligns with the context considered in these studies.

3.6 Summary and Discussion

The purpose of our study is to verify the finding of Cappelen et al. (2007) on the pluralism of fairness ideals—namely, that various fairness ideals coexist within the population. Cappelen et al. (2007) and previous experimental studies set endowments to be determined by the production stage before the dictator game so that an individual's luck and efforts are involved in the distribution problems. In this situation, three fairness ideals regarding what the fair distribution should be—strict egalitarianism, libertarianism, and liberal egalitarianism—correspond to different answers to the distribution problem. To demonstrate the existence of pluralism with respect to these three ideals, they suggest a choice model assuming that an individual incorporates one of the three fairness ideals into his or her utility function for the distribution problems. Based on this model, they attempt to estimate the population distributions of adherence to the three fairness ideals.

We introduce a pure egoist to the choice model of Cappelen et al. (2007), i.e., individuals who prefer to take all of the product. In other words, we modify the model of Cappelen et al. (2007) by assuming four types: strict egalitarians, libertarians, liberal egalitarians, and egoists. As a result, around 40% of subjects are classified as egoists, representing a substantial majority. However, notwithstanding this majority, the three fairness ideal types proposed by Cappelen et al. (2007) still coexist. Moreover, the distribution of fairness ideals also changes considerably when we use our modified models for estimation. The estimation result using Cappelen et al. (2007) model shows a higher proportion of libertarians than of liberal egalitarians. However, the result of our modified estimation models indicates that the proportion of the liberal egalitarian type is larger than other fairness ideal types, such as strict egalitarians and libertarians. Additionally, by comparing the values of the weight parameter across the fairness ideals, we find that strict egalitarians assign the highest weight to their ideal, followed by liberal egalitarians, with libertarians having the lowest weight among the three types. This result suggests that when considering fair distribution, the more factors attributable to the individual—such as individual luck and effort—are included, the lower the weight placed on the ideal tends to be.

We also assess the fitness of our modified models and Cappelen et al. (2007)'s model. In the t-test, we find no significant difference in observations between Cappelen et al. (2007)'s model and our models. However, in the variance ratio test, only our modified models are not significantly different from the observations. These results suggest that our modified models demonstrate a better fit than the model of Cappelen et al. (2007), supporting our assumption that the choice model of distribution decision should include pure egoism.

In addition, individuals with high effort in the task tend to be classified as libertarians. Conversely, those with high productivity are more likely to be classified as liberal

egalitarians and egoists than other types. The proportion of strict egalitarians and liberal egalitarians is significantly higher among women than among men, while the proportion of egoists is significantly higher among men than among women. These findings align with the results ofMiller and Ubeda (2012) and Sharma (2015), indicating that women tend to base their distribution decisions on effort rather than luck.

In conclusion, our study provides important insights into what is considered a fair distribution in society. First, individual selfishness cannot be overlooked: when given the authority to distribute the endowment, as observed in the dictator game, a significant number of participants exhibit egoistic behavior, irrespective of individual effort, luck, or other factors. Second, even when individual selfishness has a non-negligible impact on distributional issues, strict egalitarianism, libertarianism, and liberal egalitarianism exert an inviolable influence. Although the sociocultural environment influences debates on which is the fairest–strict egalitarianism, libertarianism, or liberal egalitarianism–more than half of the participants adhere to one of these three ideals without resorting to selfish choices, even when granted authority over distribution. This suggests that, in distribution decisions, individuals express distributional preferences rather than conforming strictly to the homo economicus model. Therefore, our study not only facilitates more realistic estimation by revealing the existence of individual selfishness in distribution decisions but also clarifies that pluralism in fairness ideals persists, as shown in previous studies.

Chapter 4

Heterogeneity in the Effect of Empathy on Plural Fairness Ideals

"By the imagination we place ourselves in his situation, we conceive ourselves enduring all the same torments, we enter as it were into his body, and become in some measure the same person with him, and thence form some idea of his sensations, and even feel something which, though weaker in degree, is not altogether unlike them."
"Those general rules of conduct, when they have been fixed in our mind by habitual reflection, are of great use in correcting the misrepresentations of self-love concerning what is fit and proper to be done in our particular situation."
Smith (1759/2010), The Theory of Moral Sentiments

4.1 Introduction

People who live in society are self-conscious and want to make appropriate decisions that are generally acceptable to the members of their society even when others have no room for influence on individual decision-making and monetary payoffs. In his first book, *The Theory of Moral Sentiments*, Adam Smith, the father of economics, argues that the "impartial spectator" of conscience plays an important role in individual behavior (Ashraf, Camerer, & Loewenstein, 2005).

In this study, following the idea of Haidt (2012), we examine the effect of empathy on plural fairness ideal types through dictator game experiments. Firstly, we estimate individual fairness ideal types according to model M1 in Section 3.4.3 of Chapter 3. In Chapter 3, we argue that it is difficult to represent strict egalitarianism in Cappelen et al. (2007)'s experimental setting based on a dictator game. We show that 40% of participants behave as egoists who take all the endowment, which is a rational choice in a dictator game. Moreover, 16% of participants classified as strict egalitarians, 19% as libertarians, and 25% as liberal egalitarians. Thus, in this study, we adopt the estimation model of Section 3.4.3 to examine whether promoting state empathy differently affects people with different fairness ideals.

Second, we investigate the effect of empathy on each fairness ideal type. Many studies show that the effect of empathy on fairness considerations is generally positive (Edele et al., 2013; Klimecki, Mayer, Jusyte, Scheeff, & Schönenberg, 2016; Singer, 2006; Singer et al., 2006; Urbanska, McKeown, & Taylor, 2019). If this is true, is the degree of this positive effect the same for people with different fairness ideals? Based on moral foundation theory,

Haidt (2012) shows that politically liberal people are more disturbed by others' suffering than libertarians. With clues from Haidt (2012), we try to determine whether there is a difference in the effect of empathy across individual fairness ideals using dictator game experiments. We take the concept of *state empathy*, which is a subtype of empathy that increases in specific situations because the aim of our study is to examine the differences in the effect of empathy on each fairness type by comparing conditions that promote empathy with those that do not.

Our experiments consist of two parts. In Part 1, we estimate the fairness ideal type of subjects using same experimental procedure as described in Section 3.2 of Chapter3. Additionally, we introduce Part 2, which is a modified dictator game based on the A(e) condition of Andreoni and Rao (2011)[1] to highlight state empathy and examine the effect of state empathy on subjects with different fairness ideals.

As a result, similar to previous studies, in the empathy treatment, where state empathy is promoted, the allocator's own share decreases, that is, the altruistic distribution generally increases on average. However, the direct effects vary depending on ideal type. The egoistic allocators reduce their shares the most, followed by the libertarian and liberal egalitarian allocators. In addition, we find that the gap between the ideal shares of those with libertarian or liberal egalitarian fairness ideals and their distribution decisions decreases. Moreover, the asking messages do not significantly affect the distribution decisions of any of the allocators when the decisions made before and after reading the messages from the receiver are compared. Thus, we conclude that in a situation where each individual's efforts and luck determine the endowment to be distributed, state empathy reduces selfishness in different ways for each type. For those who make an egoistic choice when they have the power to distribute, that is, for those who do not have a specific fairness ideal, state empathy directly increases altruistic distribution; however, for those who have fairness ideals, such as libertarians and liberal egalitarians, it seems to reduce selfishness and put more weight on fairness ideals.

This chapter proceeds as follows. First of all, Section 4.2, we describe the experimental design based on Cappelen et al. (2007), Almås et al. (2010), and Andreoni and Rao (2011). Section 4.3 presents the results, and Section 4.4 gives conclusions and a discussion.

4.2 The Experiments

Our experiments of this study consist of two parts. In Part 1, described in Section 3.2 of Chapter3, we determine the type of subjects. In Part 2, a dictator game with two treatments is used to examine the effect of state empathy. Part 2 has two treatments, namely, the empathy treatment, which is based on the A(e) condition of Experiment 2 in Andreoni and Rao (2011), and the baseline treatment, in which the roles are assigned first and only allocators perform distribution decisions; we explain the details below. All the subjects participated in only one of the two treatments in Part 2. When the subjects participate in Part 1, they are not provided with any information on Part 2. When Part 1 is finished, we inform them that an additional part will be conducted; then, after we obtain agreement from the subjects for Part 2, we provide the instruction for Part 2. In this section, we solely present the experimental procedures of Part 2[2].

After finishing Part 1, we inform all participants to perform an additional Part and obtain their agreement. Next, we distribute the new instructions for Part 2 and read them aloud via the computer. The subjects participate in only one treatment, that is, either

[1] We will explain the details of Andreoni and Rao (2011) in Section 1.2.3 in the Chapter 1.

[2] For Part 1, see Section 3.2 in Chapter 3.

Figure 4.1: A Screenshot of Stage 1 in the Empathy Treatment

the baseline or the empathy treatment. In Part 2, the endowment to be distributed by the allocator is taken from the data in Part 1. The subjects are randomly paired at the beginning of Part 2; then, the total product of group i ($X_i = x_i + x_j$) is calculated as the sum of each member's output ($x_i = a_i q_i$, $x_j = a_j q_j$) based on each individual's efforts (q_i, q_j) and productivity (a_i, a_j), which are taken from the production stage of Part 1. When the allocator makes a decision, the following information is provided: the points earned from the task by the focal participant and his or her partner (q_i, q_j), the rate of return (a_i, a_j), the output (x_i, x_j), and the group's total product ($X_i = x_i + x_j$). The subject who is given the role of allocator makes a distribution decision on the group's total product X_i.

The first treatment, that is, the empathy treatment, which is based on the A(e) condition in Andreoni and Rao (2011)'s experiments, consists of two stages. Since the roles are not assigned in Stage 1, the subjects make decisions without knowing their roles. Stage 1 is the situation of putting oneself in another's shoes. All the subjects make distribution decisions and write explanation messages to a receiver as an allocator and ask distribution questions and write asking messages to the allocator as a receiver. A screenshot of Stage 1 in the empathy treatment is shown in Figure 4.1. Hereafter, we call Stage 1 of the empathy treatment the Role Uncertainty (RU) condition. When all the decisions and the writing are completed, the roles are assigned at the beginning of Stage 2. The subjects in the role of the allocator receive their partners' distribution ask and messages, and the subjects in the role of the receiver receive distribution decisions and explanation messages from the allocator. At this time, the allocators make a final decision after reading their partners' distribution ask and messages, and the receivers cannot make any decisions, as is the case in the baseline, and wait for the allocators' final decisions. Hereafter, we call Stage 2 of the empathy treatment the Asking (A) condition. In Part 2, the final decision of the allocator is given as payoffs, and all subjects are informed of all the processes of Part 2 and the information provided to the allocator in the instructions.

Second, we construct the baseline treatment as a control treatment. In this treatment, subjects are randomly paired in groups and the roles of the allocator and the receiver are assigned to each member. The endowment is given to only the allocator, and he or she is

Table 4.1: Summary of Sessions

Session	Treatment	Time limit for production (seconds)	Average efforts (points)	Standard deviation	Number of subjects	Total subjects
1	Baseline	120	15.7	5.47	30	212
3		120	24.3	10.29	20	
5		150	15.3	6.54	18	
8		150	17.5	8.08	26	
9		150	19.8	8.00	20	
11		150	15.2	8.10	26	
12		150	16.6	7.64	20	
14		150	16.2	7.69	26	
15		150	17.7	6.46	26	
2	Empathy	120	8.9	3.38	30	150
4		150	14.4	6.21	18	
6		150	18.4	4.88	26	
7		150	22.1	12.40	20	
10		150	18.2	7.11	28	
13		150	15.8	6.92	28	
Total			16.8	8.00		362

free to distribute it to the partner in any way, including keeping it all for him or herself. The receiver does not have any effect on the allocator's distribution decision, and his or her payoffs are determined by the allocator's decision.

The experiments were conducted at Waseda University from the fall of 2019 to the spring of 2023. We ran fifteen sessions with 18 to 30 subjects in each session, and all 362 participants were students from Waseda University (Table 4.1). The experiments were computerized with z-Tree (Fischbacher, 2007). Each session lasted about 70 minutes, and the show-up fee was ¥1000 (≈ \$7.9); average earnings were ¥1510 (≈ \$11.9)[3].

4.3 Results

4.3.1 The Type Estimation

We use the choice model of Cappelen et al. (2007) and the estimation model M1 in Section 3.4.3 which add the egoistic ideal to the choice model of Cappelen et al. (2007). We assume that there are four fairness ideals. First, the strict egalitarian (SE) ideal is that distribution should be equal regardless of individual's luck and efforts ($m^{SE(i)}$). Second, the libertarian (L) ideal is that each individual has the right to possess his/her own output because both an individual's productivity and an individual's efforts are under that individual ($m^{L(i)}$). Third, the liberal egalitarian (LE) ideal claims that each individual is responsible for only his/her choice, that is, efforts, and that the effects of luck should be excluded ($m^{LE(i)}$). Fourth, the egoistic (EGO) ideal claims that taking all the endowments is a reasonable choice when possible ($m^{EGO(i)}$). The fair shares according to each fairness ideal are as

[3]We convert Japanese yen into US dollars at the exchange rate of \$1=¥127, which was the average rate at the time the experiments were conducted, that is, from the fall of 2019 to the spring of 2023.

follows:

$$\begin{aligned} m^{SE(i)} &= \frac{X_i}{2} \\ m^{L(i)} &= a_i q_i \\ m^{LE(i)} &= \frac{q_i}{q_i + q_j} X_i \\ m^{EGO(i)} &= X_i \end{aligned}$$

where q_i is the level of effort (in our experiment, the total points from the slider task), a_i is productivity (in our experiment, the rate of return, which can be high or low), and X_i is the total product of the group of i.

We adopt the finite mixture model in Moffatt (2015), and the approach of Moffatt (2015) is as follows: First, the model of the four fairness ideals is constructed, and a label is assigned to each. Second, a parametric model is specified for the behavior of each ideal. Third, the parameters of these three models are estimated together, along with the "mixing proportion" – the proportion of the subjects who follow each ideal. Finally, the posterior probability of each subject following each ideal is determined.[4] Moreover, we assume that all three errors have the same variance as follows: $V(\epsilon_{1,i}) = V(\epsilon_{2,i}) = V(\epsilon_{3,i}) = V(\epsilon_{4,i}) = \sigma^2$.

The mixing proportions of each ideal are p_1, p_2, p_3, and p_4, and the likelihood contribution of each subject i is as follows:

$$\begin{aligned} L_i = p_1 \frac{1}{\sigma}\phi\left(\frac{y_i - \frac{X_i}{2} - \frac{X_i}{\beta}}{\sigma}\right) + p_2 \frac{1}{\sigma}\phi\left(\frac{y_i - a_i q_i - \frac{X_i}{\beta}}{\sigma}\right) \\ + p_3 \frac{1}{\sigma}\phi\left(\frac{y_i - \frac{q_i}{q_i + q_j} X_i - \frac{X_i}{\beta}}{\sigma}\right) + (1 - p_1 - p_2 - p_3)\frac{1}{\sigma}\phi\left(\frac{y_i - X_i}{\sigma}\right) \end{aligned}$$

where ϕ is probability density function of normal distribution. The parameters to be estimated are β, σ, p_1, p_2, and p_3, and we conduct log-likelihood estimation.

Moreover, according to Moffatt (2015), the posterior probability that each subject follows each fairness ideal can be calculated using the estimated parameters from the mixture model. The equation for calculating the posterior probability of following each fairness ideal by using the estimated values of the parameters is as follows:

1) Strict egalitarian (SE)

$$P(i = \text{SE}|y_{i1}, ..., y_{iT}) = \frac{p_1 \prod_{t=1}^{T} \frac{1}{\sigma}\phi\left(\frac{y_i - \frac{X_i}{2} - \frac{X_i}{\beta}}{\sigma}\right)}{L_i}$$

2) Libertarian (L)

$$P(i = \text{L}|y_{i1}, ..., y_{iT}) = \frac{p_2 \prod_{t=1}^{T} \frac{1}{\sigma}\phi\left(\frac{y_i - a_i q_i - \frac{X_i}{\beta}}{\sigma}\right)}{L_i}$$

[4]However, any given subject's fairness ideal cannot be identified with certainty.

Table 4.2: Type Distribution

	Estimation results	Posterior type distribution
Proportion of type SE	0.1900	16.0%
	(0.0139)	(58 subjects)
Proportion of type L	0.1906	19.1%
	(0.0130)	(69 subjects)
Proportion of type LE	0.2511	25.4%
	(0.0140)	(92 subjects)
Proportion of type EGO	0.3683	39.5%
	(0.0109)	(143 subjects)
β	9.2533	
	(0.2958)	
σ	11.3246	
	(0.0993)	
Log Likelihood	-31166.482	

Note: In the second column, standard errors are given in parentheses. SE: strict egalitarian, L: libertarian, LE: liberal egalitarian, EGO: egoist.

3) Liberal egalitarian (LE)

$$P(i = \text{LE}|y_{i1}, ..., y_{iT}) = \frac{(1 - p_1 - p_2) \prod_{t=1}^{T} \frac{1}{\sigma} \phi \left(\frac{y_i - \frac{q_i}{q_i + q_j} X_i - \frac{X_i}{\beta}}{\sigma} \right)}{L_i}$$

4) Egoist (EGO)

$$P(i = \text{EGO}|y_{i1}, ..., y_{iT}) = \frac{p_4 \prod_{t=1}^{T} \frac{1}{\sigma} \phi \left(\frac{y_i - X_i}{\sigma} \right)}{L_i}$$

This calculation is based on Bayesian rules. Using the six decisions made by each subject in the distribution stage, we calculate the posterior probability of following each fairness ideal, and for each subject, the fairness ideal with the highest probability value is that subject's ideal type.

According to the estimation model above, the estimated proportion of each type, the weight given to fairness (β), and the number of subjects of each type based on the posterior type probability calculations are shown in Table 4.2.

In Table 4.2, the second column shows the estimated proportion of each type and the weight given to fairness (β). As a result, the estimated type distribution is as follows: 19.00 percent correspond to the strict egalitarian type, 19.06 percent to the libertarian type, 25.11 percent to the liberal egalitarian type, and 36.83 percent to the egoistic type.

In addition, as shown in the third column of Table 4.2, the distribution derived by posterior calculation is not much different from the distribution estimated by log-likelihood estimation in the first column of Table 4.2: 16.0 percent correspond to the strict egalitarian type, 19.1 percent to the libertarian type, 25.4 percent to the liberal egalitarian type, and 39.5 percent to the egoist type. Therefore, the types of subjects derived by posterior calculation are valid. In the following sections, based on the posterior fairness ideal types

Table 4.3: Basic Statistics of Allocator's Share in Each Treatment

Baseline : Allocator's share	Total	SE	L	LE	EGO
Obs.	106	18	29	18	41
Avg.	0.7678	0.6562	0.6511	0.5819	0.9810
SD	0.2362	0.2028	0.1962	0.2254	0.0532
Min	0	0.4861	0.3438	0	0.7143
Max	1	1	1	0.9889	1

Empathy (RU): Allocator's share	Total	SE	L	LE	EGO
Obs.	150	23	24	44	59
Avg.	0.6495	0.5483	0.5535	0.5853	0.7758
SD	0.2190	0.1289	0.1190	0.1527	0.2584
Min	0.1712	0.3765	0.3485	0.2500	0.1712
Max	1	1	1	0.8661	1

Empathy (A): Allocator's share	Total	SE	L	LE	EGO
Obs.	75	10	14	20	31
Avg.	0.6553	0.5773	0.5969	0.5846	0.7524
SD	0.2304	0.1601	0.1724	0.1613	0.2779
Min	0	0.4167	0.4322	0.3691	0
Max	1	1	1	1	1

Note: Empathy (RU) is Stage 1 of the empathy treatment, in which the roles are uncertain, and Empathy (A) is Stage 2 of the empathy treatment, in which the subjects who are assigned the role of the allocator can read the ask of the receiver.

of the subjects, we examine whether the effect of empathy on each fairness ideal type differs.

4.3.2 State Empathy Encourages Fairness Considerations

Table 4.3 shows the basic statistics of the allocator's share in the baseline treatment (Baseline, the upper table in Table 4.3), in Stage 1 of the empathy treatment (Empathy (RU), the middle table in Table 4.3), and in Stage 2 of the empathy treatment (Empathy (A), the lower table in Table 4.3).

First, state empathy, which is promoted in the empathy treatment, is effective in terms of increasing the altruism of the distribution of allocators except liberal egalitarians. Figure 4.2 displays the results regarding the average share of allocators in the baseline treatment; the RU condition of the empathy treatment, in which decisions are made before asking messages are read; and the A condition of the empathy treatment, in which decisions are made after the asking messages are read. The figure shows that the overall average share taken by allocators decreased from the baseline treatment to the empathy treatment in the case of type SE, L, and EGO; moreover, the difference is statistically significant (t-text, $p = 0.0223$ for type SE; $p = 0.0189$ for type L; $p = 0.0000$ for type

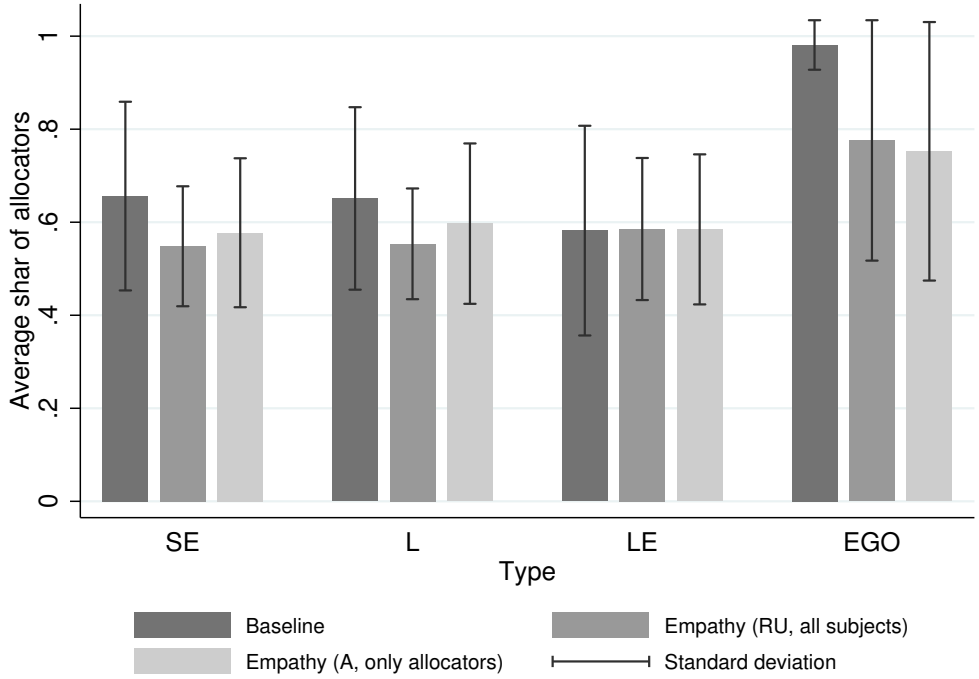


Figure 4.2: Average Share of the Allocators before Role Assignment and after Reading Asking Messages.

EGO).

More specifically, the effect of state empathy differs between the fairness ideal types. In Figure 4.2, comparisons of the allocators' average shares between the baseline treatment and the empathy treatment (RU) are shown by type. First, as shown in the graphs for type EGO, who do not have a specific fairness ideal and behave the most selfishly, reduce their share the most. On average, they take 98.10% of the total product in the baseline treatment but significantly reduce their share to an average of 77.58% in the empathy treatment (RU) (t-test, $p = 0.0000$). Second, type SE and L subjects also reduce their share from about 65% in the baseline treatment to 55% in the empathy treatment (RU), as shown in the graphs for type SE and L, and their average share in the empathy treatment (RU) is significantly lower than that in the baseline treatment at the 5% level (t-test, $p = 0.0223$ and $p = 0.0189$, respectively). Third, as shown in the graphs for type LE, type LE subjects show little change in their share between the baseline and the empathy treatment (RU), which amount to 58.19% and 58.53% of the total, respectively. Their average share in the empathy treatment (RU) is not significantly different from that in the baseline treatment (t-test, $p = 0.5276$).

Since our experiments allow the endowment to be distributed on the basis of each subject's efforts and luck, the overall decrease in the empathy treatment is smaller than that reflected in the results of Andreoni and Rao (2011). However, our results are also in line with Andreoni and Rao (2011)'s results, which show that the promotion of state empathy decreases the allocator's share and increases the shares of others in distribution decisions. Furthermore, in particular, we find that the share-reducing effect of state empathy promotion manifests differently in people with different fairness ideals.

In addition, the percentage of subjects who take all the total product, that is, those who are zero-offering, decreases as shown in Table 4.3 and Figure 4.3. In general, in the dictator

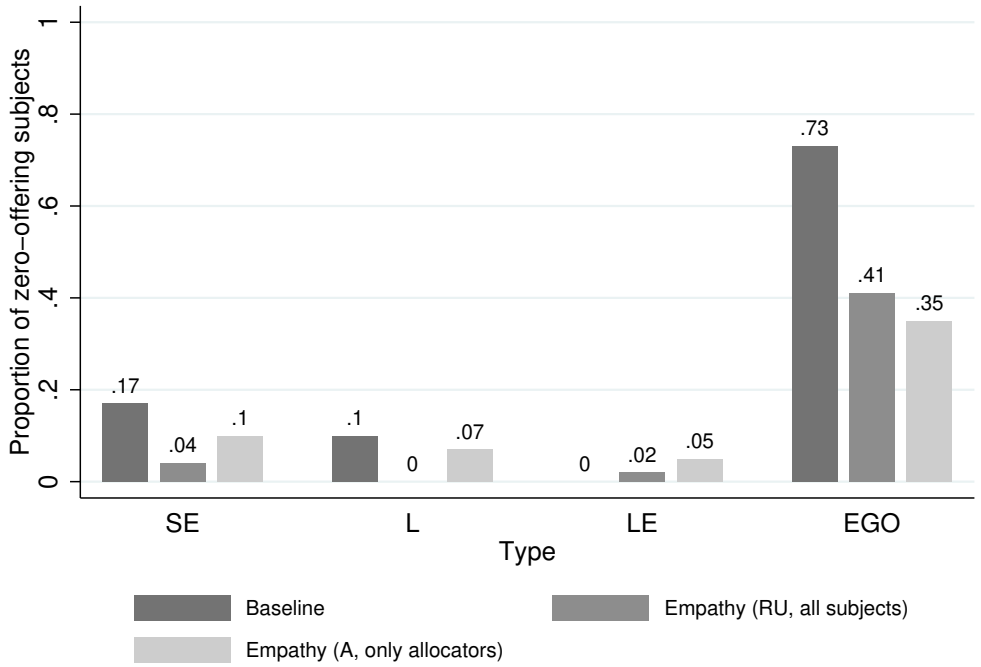


Figure 4.3: Proportion of Zero-Offering Subjects.

game, about 36% of the allocators offer nothing to their partners Engel (2011); moreover, in our experiments, 33.96% of the allocators offer nothing in the baseline treatment. However, in the empathy treatment, the percentage of zero-offering subjects significantly decreases to 17.33% under the RU condition and to 19.44% under the A condition.

However, the decrease in the percentage of zero-offering subjects in the empathy treatment differs across the fairness ideal types. As shown in the graphs for type EGO in Figure 4.3, as the ideal of this type involves taking all the product, 73% of the type EGO subjects in the baseline treatment make zero-offer. However, this percentage decreases remarkably in the empathy treatment to 41% of the subjects, implying that state empathy encourages type EGO individuals to behave more altruistically. Compared to type EGO subjects, type SE, type L and LE subjects show a lower percentage of zero-offering subjects in the baseline game, that is, 17% for type SE, 10% for type L, and 0% for type LE. Since type SE, type L and type LE subjects have fairness ideals that involve distributing endowments according to individual luck and effort, the proportion of these subjects making no offer is low even when they are assigned the role of allocator and they have authority over distribution.

Second, the asking message seems to have little effect on the decision-making of the allocators. We examined the average allocator's share before and after reading the receiver's message in the empathy treatment by type, and compared it to the average share of the allocators in the RU condition with the A condition of the empathy treatment (Figure 4.2). Our results show that the average share of the allocators is not significantly different for each type (t-test: type SE, $p = 0.5841$; type L, $p = 0.3646$; type LE, $p = 0.9859$; type EGO, $p = 0.6917$). This suggests that asking messages from receivers are unlikely to change the distribution decisions of the allocators. Interestingly, we observed a slight increase in the number of people who make no offer after reading the asking message in the empathy treatment for type SE (10%), for type L (7%), and type LE (5%), as shown in

Figure 4.3. This result suggests that some people with fairness ideals may respond to the specific contents of the messages by taking all the product as an expression of disagreement with those messages Bruttel and Stolley (2018).

In summary, the allocator's distribution decision is affected by the promotion of state empathy, and this effect differs by type. Type EGO is most affected by state empathy, followed by type SE and type L, and type LE is rarely affected. Moreover, Drange Hole (2011) shows that communication rather than the content of communication has a positive effect on the distribution decision, and this evidence may support our results. In other words, promoting state empathy decreases the allocator's share except tyep LE; however, the content of asking messages may not have a significant effect on an allocator's distribution decision because there is no difference between the periods before and after receivers' messages are read for any type.

Next, we examine the effect of state empathy by type in more detail. In the following sections, based on our results that there is no significant difference in the distribution decision before and after the asking message from the receiver is read, our analyses are carried out with the data from the role uncertainty (RU) condition in the empathy treatment, in which all the participants make distribution decisions as allocators before role assignment.

4.3.3 The Egoistic Allocator is the Most Empathetic

In Table 4.3 and Figure 4.2, the difference in the average allocators' share between the baseline treatment and the empathy treatment (RU) can be compared. Overall, there is a 11.83% decrease in the average share. However, when considering each type separately, type EGO subjects exhibit the largest reduction in their shares, with an average decrease of 20.52% in the empathy treatment (RU). This is followed by an average decrease of 10.79% for type SE and 9.76% for type L.

This result implies that there are significant differences in the effect of state empathy by type. To clarify this difference between types, we perform Tobit regressions for each type using allocator i's share per his/her distribution decision as a dependent variable, and the results are shown in Table 4.4. The independent variables are the treatment dummy variable, with 0 for the baseline treatment and 1 for the empathy treatment (RU); i's productivity ($a_i \in \{2, 4\}$), which is randomly assigned; i's efforts (q_i) from the real effort task; and the difference in efforts between allocator i and receiver j ($q_i - q_j$). We use normalized q_i and $q_i - q_j$ for all Tobit regression analyses because i's efforts (q_i) are represented by a value between 0 and 48 and the range of these values is much greater than that of the allocator's share, which is a dependent variable with a value between 0 and 1. This makes the coefficient of q_i too small, so all q_i and $q_i - q_j$ are normalized to a value between 0 and 1[5]. In addition, we include the variable $q_i - q_j$, which is the difference in effort between the allocator and his/her receiver, because there is a possibility that task performance is considered a competition between the subjects. In this case, the

[5]We use the min-max normalization as follows:

$$\text{norm\_}q_i = \frac{q_i - \min_{k \in I} q_k}{\max_{k \in I} q_k - \min_{k \in I} q_k} = \frac{q_i - 0}{48 - 0} = \frac{q_i}{48}$$

$$\begin{aligned}\text{norm\_}(q_i - q_j) &= \frac{(q_i - q_j) - \min_{k \neq l}(q_k - q_l)}{\max_{k \neq l}(q_k - q_l) - \min_{k \neq l}(q_k - q_l)} \\ &= \frac{(q_i - q_j) - (-27)}{27 - (-27)} = \frac{(q_i - q_j) + 27}{54}.\end{aligned}$$

where I is the set of subjects.

subjects would make decisions on the basis of a comparison of their efforts with those of their partners as if they were determining who was the winner. The two variables, that is, the treatment dummy and the productivity variable a_i, are used in Models (1), (2), and (3) as independent variables. In addition, i's efforts, q_i, are used in Model (1), the difference in efforts $q_i - q_j$ instead of i's efforts q_i is used in Model (2), and both are used in Model (3) as independent variables.

First, the coefficients for type EGO are the largest and show high significance only for the treatment dummy variables. This result confirms that the distribution decision of type EGO is more strongly affected by state empathy than the decisions of the other three types. The results regarding type EGO in Table 4.4 have coefficients of -0.4119 in Model (1), -0.4216 in Model (2), and -0.4310 in Model (3) at the 0.1% level of significance for the treatment dummy variables in the three models. Since type EGO subjects tend to take all the product if possible and do not have a specific fairness ideal, it is reasonable that there are no significant variables other than the treatment dummy variables.

Second, types SE and L are also affected by state empathy, but the magnitude and significance of the effect are smaller than that of type EGO. In the case of type SE in Table 4.4, the treatment dummy variables have coefficients of -0.1236 in Model (1), -0.1206 in Model (2), and -0.1196 in Model (3) at the 5% level of significance. In the case of type L, the treatment effect is slightly stronger than that for type SE when we include the variable $q_i - q_j$ in Models (2) and (3), because the magnitude of treatment dummy variables has a coefficient of -0.1016 in Model (1) at the 5% level, a coefficient of -0.1302 in Model (2) at the 1% level, and a coefficient of -0.1335 in Model (3) at the 0.1% level. All the significant coefficients of the treatment dummy variables are significantly smaller than the those corresponding to type EGO (Wald test: SE vs. EGO, $p = 0.0035$ and L vs. EGO, $p = 0.0004$ in Model (1); SE vs. EGO, $p = 0.0025$ and L vs. EGO, $p = 0.0008$ in Model (2); SE vs. EGO, $p = 0.0018$ and L vs. EGO, $p = 0.0006$ in Model (3)).

Third, the productivity variables a_i are significant in the case of types L and EGO in all the models. For type L, this result confirms that focal subjects have the libertarian fairness ideal, asserting that randomly assigned productivity should be considered for distribution along with individual efforts. The productivity variables have significant coefficients of 0.1002 in Model (1), 0.1001 in Model (2), and 0.1081 in Model (3) at the 0.1% level. Type EGO also shows significance in the productivity variables, but only at the 5% level in all three models. This suggests the possibility that the EGO type not only engages in egoistic distributions by attributing the lucky factor (i.e., the role of the allocator) to themselves but also regards another luck-driven factor, productivity, as their own, thereby adjusting their share of the distribution accordingly.

Fourthly, only type LE remains unaffected by state empathy. The results in Table 4.4 indicate that type LE subjects lack a significant coefficient for the treatment dummy variables. Furthermore, in all models, variables related to efforts, namely one's own efforts (q_i) and the differences in individuals' efforts ($q_i - q_j$), prove significant for type LE. In Model (1), a significant value of 0.5116 is observed for q_i, while Model (2) records a significant value of 0.7825 for the difference in efforts ($q_i - q_j$) at the 0.1% level. Interestingly, Model (3) incorporates both one's own efforts (q_i) and the difference in efforts ($q_i - q_j$) as independent variables, with both being significant–0.2211 at the 5% level and 0.6678 at the 0.1% level. These results suggest that type LE subjects base their distribution decisions only on individual efforts, aligning with their liberal egalitarian fairness ideals, remaining unaffected by state empathy or other factors. Moreover, according to the results of Model (3), they distribute based on both individual efforts and the differences in efforts, with the level of efforts itself exerting a weaker influence than the differences.

Table 4.4: Tobit Regressions

Dependent Variable:	(1)				(2)				(3)			
Allocator i's share	SE	L	LE	EGO	SE	L	LE	EGO	SE	L	LE	EGO
Treatment (0: Base, 1: Empathy)	-0.1236^{*} (0.0565)	-0.1016^{*} (0.0392)	0.0071 (0.0441)	-0.4119^{***} (0.0859)	-0.1206^{*} (0.0552)	-0.1302^{**} (0.0363)	0.0087 (0.0367)	-0.4216^{***} (0.0854)	-0.1196^{*} (0.0564)	-0.1335^{***} (0.0352)	0.0063 (0.0357)	-0.4310^{***} (0.0856)
Productivity, a_i	0.0297 (0.0288)	0.1002^{***} (0.0207)	-0.0183 (0.0216)	0.0952^{*} (0.0394)	0.0350 (0.0293)	$.1001^{***}$ (0.0182)	0.0051 (0.0185)	0.0888^{*} (0.0389)	0.0352 (0.0294)	0.1081^{***} (0.0182)	0.0058 (0.0180)	0.0874^{*} (0.0387)
Efforts, q_i	-0.1100 (0.1473)	0.5127^{**} (0.1746)	0.5116^{***} (0.1215)	-0.0735 (0.2422)					0.0196 (0.2119)	0.2936 (0.1625)	0.2211^{*} (0.1086)	-0.2937 (0.2737)
Difference in efforts, $q_i - q_j$					-0.1717 (0.1526)	0.5251^{***} (0.1136)	0.7825^{***} (0.1113)	0.2703 (0.2175)	-0.1864 (0.2209)	0.4511^{***} (0.1176)	0.6678^{***} (0.1220)	0.4022 (0.2473)
Constant	0.6259^{***} (0.0989)	0.1885 (0.1007)	0.4681^{***} (0.0925)	1.0252^{***} (0.1505)	0.6564^{***} (0.1047)	0.1234 (0.0854)	0.1888 (0.0953)	0.8893^{***} (0.1525)	0.6561^{***} (0.1048)	0.0325 (0.0968)	0.1681 (0.0933)	0.9310^{***} (0.1565)
Observations	41	53	62	100	41	53	62	100	41	53	62	100

Note: Variables q_i and $q_i - q_j$ are normalized. Standard errors are given in parentheses. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Fifth, individuals classified as Type SE consistently distribute nearly half of the total product without considering productivity and effort. This is supported by the sustained significance of the constant across all three models at the 0.1% level for Type SE, with values of 0.6259 in Model (1), 0.6564 in Model (2), and 0.6561 in Model (3). These findings highlight that individuals of Type SE, even in their role as allocators, tend to distribute the total product in close alignment with their fairness ideal. Notably, under the empathy treatment, state empathy has a valid effect, as indicated in the second result. That is, state empathy offsets their distribution, which is slightly more selfish than half in the baseline, and makes them adhere strongly to their fairness ideal, strict egalitarianism.

At last, when type L subjects incorporate individual efforts into their distribution decisions, they tend to perceive individuals' performance in the task as a competition. Model (3) results in Table 4.4 reveal that they adjust their distribution based on the difference in efforts between themselves and others, with the coefficient of the difference in efforts being 0.4511 at the 0.1% significance level in Model (3). These findings confirm that type L subjects consider both randomly assigned productivity and individual efforts in their distribution decisions, with a slightly greater emphasis on the difference in individual efforts in their decision-making process.

In summary, we confirm that the effect of state empathy differs by type. type EGO subjects react to the promotion of state empathy and decrease their shares the most among the four types, and type LE subjects are hardly affected by the promotion of state empathy. In other words, there is little or no effect of state empathy for people who have fairness ideals, such as strict egalitarianism, libertarianism or liberal egalitarianism; this result differs from the results of previous studies, which show that state empathy increases fairness considerations in general. In addition, we find that when people consider individual efforts on a task in their distribution decisions, libertarians and liberal egalitarians may tend to compare efforts and perceive effort as a competition.

4.3.4 State Empathy Reduces Selfishness

Although promoting state empathy has the strongest influence on type EGO allocators, as shown so far, their average share is still the highest (Table 4.3). We perform a t-test to compare the average shares of the allocators by type in the baseline and empathy treatments. As a result, in the baseline treatment, the average share of the type EGO allocators is the highest. There are significant differences between type EGO and the other three types in the baseline treatment. However, when we compare among the three types—type SE, type L, and type LE—there is no significant difference (t-test: SE vs. L, $p = 0.4661$; SE vs. LE, $p = 0.1530$; L vs. LE, $p = 0.1364$; SE vs. EGO, $p = 0.0000$; L vs. EGO, $p = 0.0000$; LE vs. EGO, $p = 0.0000$).

Moreover, in the empathy treatment (RU), only type EGO takes a significantly higher share than the other types (t-test: EGO vs. SE, $p = 0.0001$; EGO vs. L, $p = 0.0001$; EGO vs. LE, $p = 0.0000$), and the other three types show no significant differences in the empathy treatment (t-test: SE vs. L, $p = 0.4428$; SE vs. LE, $p = 0.1622$; L vs. LE, $p = 0.1899$). That is, even though promoting state empathy has the effect of decreasing selfish distribution most effectively for type EGO, the distribution decisions based on individual fairness ideals are still more altruistic than the distribution decisions of egoists who are influenced by state empathy.

Figure 4.4 suggests another possible way in which state empathy reduces selfishness, especially in the case of types SE and L. It shows the average difference by comparing individual i's distribution decision with his/her ideal share. From the results of this figure, promoting state empathy may reduce selfishness in different ways for type EGO and types

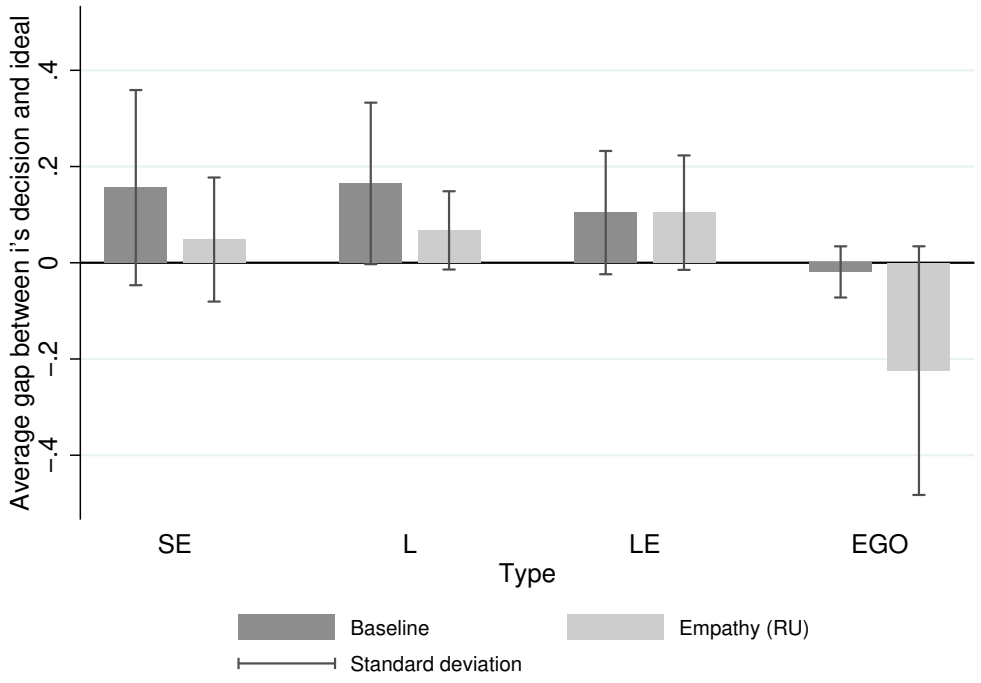


Figure 4.4: Average Gap between the Ideal Distribution and the Distribution Decision.

SE and L as follows: (1) for type EGO, state empathy induces an altruistic distribution, and (2) for type SE and L, state empathy may reduce individual selfishness and induce a distribution closer to one's fairness ideals.

First, for type EGO, as their ideal distribution is to take everything, the difference between their distribution decision and ideal distribution is, on average, negative, as shown in Figure 4.4 In the baseline treatment, type EGO subjects distribute, on average, 1.9% more to their partners than their ideal distribution, where they take all the total product. In the empathy treatments (RU), this difference widens further, as type EGO subjects take 22.4% less than their ideal on average. This average difference between their ideals and decisions is statistically significant at the 0.1% level (t-test, $p = 0.0000$). Therefore, Figure 4.4 reconfirms that state empathy promotes altruistic distributions among type EGO individuals.

Second, for type SE and type L, state empathy has less impact on directly reducing allocators' shares than it does for type EGO, but it may decrease selfishness in a different way. In particular, for those who hold strict egalitarianism or libertarianism as a fairness ideal, state empathy seems to control individual selfishness and bring decisions closer to one's fairness ideals, as the results in Figure 4.4 show. In both the baseline treatment and the empathy treatment, type SE and L subjects always take a slightly larger share than their ideal share. Type SE subjects take 15.6% more than their strict egalitarian ideal share on average, and type L subjects take 16.5% more than their libertarian ideal share on average in the baseline treatment. By using the choice model of Cappelen et al. (2007) presented in Equation 3.1 of Section 3.3, individual i's optimal offer for him/herself is the sum of i's ideal share, $mk(i)$, and selfishness, X_i/β_i. The results in Figure 4.4 show that type SE and type L subjects take more than their ideal shares according to their selfishness, especially in the baseline treatment. However, in the empathy treatment, this gap is significantly reduced, as shown in Figure 4.4. Type SE subjects, on average,

take only 4.8% more than their ideal share; this difference between the baseline and the empathy treatment is significant at the 5% level (t-test, $p = 0.0223$). In the case of type L, they take only 6.7% more than their ideal share on average, and this difference significantly decreases in the empathy treatment at the 1% level (t-test, $p = 0.0060$). The reduction in an individual's own share in the empathy treatment of type SE and type L is similar. In other words, from the results in Figure 4.4, we find that type SE and type L subjects tend to keep their distributions closer to their fairness ideals in the empathy treatment. Conversely, type LE subjects do not change their distribution decision between the baseline treatment and empathy treatment (t-test, $p = 0.4987$), and it implies that the liberal egalitarian ideal has the strongest driving force between fairness ideals.

These results provide clues that state empathy may affect people with fairness ideals, such as type SE and type L individuals, in a different way. In other words, when state empathy is promoted, their degree of selfishness may decrease, bringing their decisions closer to their fairness ideals. This result suggests the possibility that state empathy plays a role in inducing people to act in a way that more accurately reflects their fairness ideals by increasing the individual weight given to fairness, β_i in the choice model of Cappelen et al. (2007). Therefore, state empathy affects people who act mainly with selfish motives, such as type EGO individuals, by inducing altruistic behavior, but it also sometimes plays a role by reducing individual selfishness and making people follow their fairness ideals more closely.

4.3.5 Messages Affect Egoists the Most

In this section, we analyze the content of messages written by subjects in the empathy treatment. Figures 4.5 and 4.6 depict the distribution of message content during Stage 1 of the empathy treatment. Subjects assumed both allocator and receiver roles in this stage.[6] The key findings are outlined below.

First, asking messages, in general, exhibit more content than explain messages. In explain messages, individuals briefly mention their distribution criteria (depicted in the three bar graphs on the left), offer a simple greeting, or not to provide a meaningful message. In asking messages, subjects simultaneously write about fairness, offer a greeting, and appeal to the partner's power, etc. The multiple content in these messages becomes apparent when comparing Figure 4.5 with Figure 4.6, where the sum of subjects writing each content exceeds the number of subjects for each type in the asking messages.

Second, individuals who adhere to one of three the fairness ideals are more inclined to write messages referencing criteria for fair distribution. In both explain messages (Figure 4.5) and asking messages (Figure 4.6), the three fairness ideal types—strict egalitarians, libertarians, and liberal egalitarians—exhibit frequencies biased toward the three leftmost bar graphs, which represent contents of fair distribution. On the other hand, egoists did not provide meaningful messages as allocators, and their messages as receivers were scattered across multiple content types, making it hard to discern any particular tendency.

Third, each fairness ideal type tended to include content in their messages consistent with their fairness ideals. The majority of both explain and asking messages referenced factors aligned with each fairness ideal. However, exceptions existed, particularly for strict egalitarians and liberal egalitarians. Strict egalitarians had an equal number of subjects emphasizing effort and equality in their explain messages. Similarly, while liberal

[6]For categorizing message content, we employed Andreoni and Rao (2011)'s method, adding it with three factors of our three fairness ideals. Each message's content reflects a combination of opinions from three reviewers.

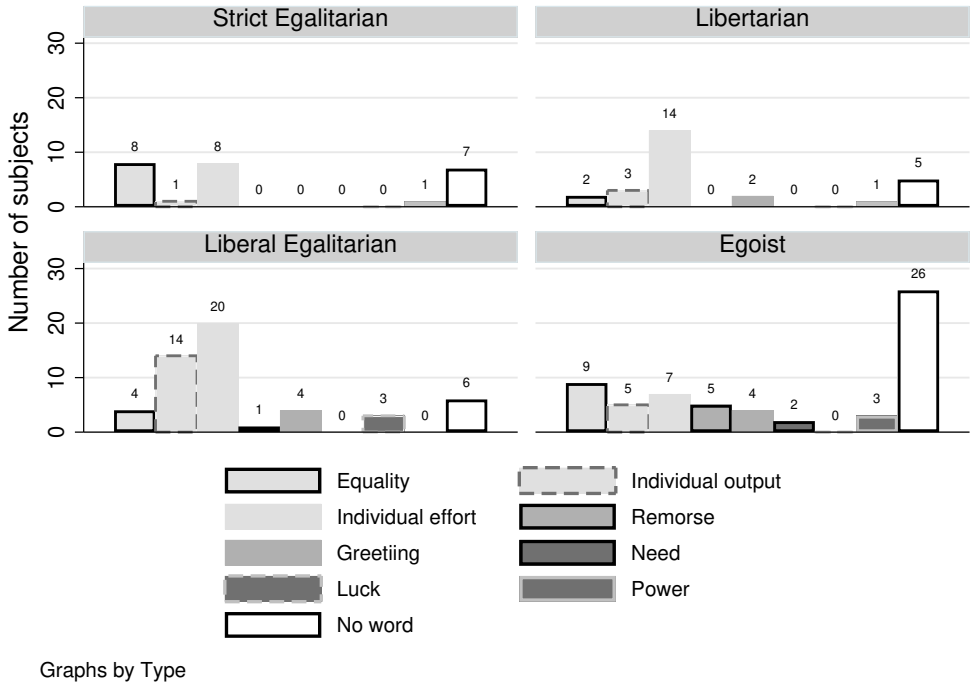


Figure 4.5: Distribution of Explain Message contents.

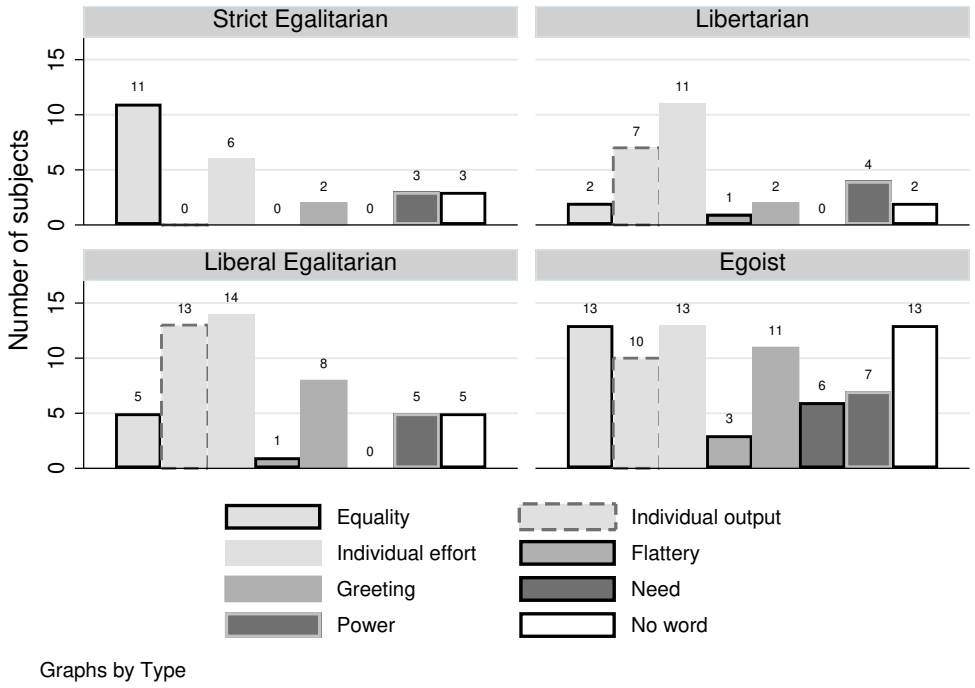


Figure 4.6: Distribution of Asking Message contents.

egalitarians were more likely to mention effort, some subjects included individual output in messages that encompassed both luck and effort.

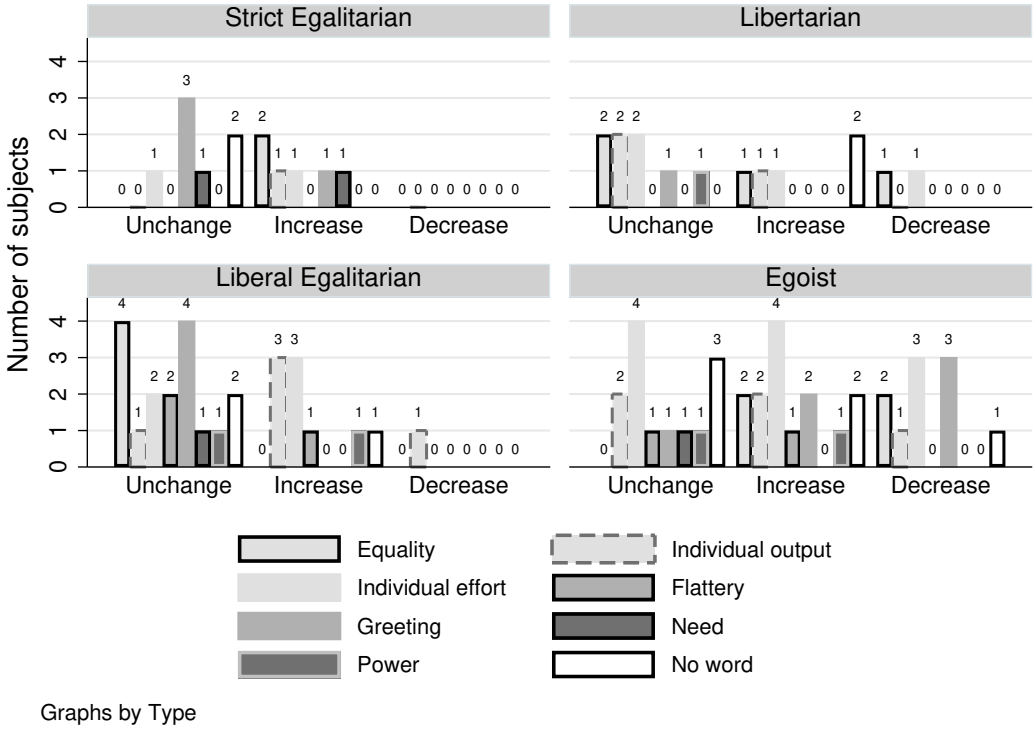


Figure 4.7: Allocator's Reaction to Message

Two potential explanations for this third result can be possible: firstly, messages in our experiments are cheap talk, preventing individuals from enforcing their fairness ideals. Additionally, due to the unknown of others' fairness ideals, there exists a possibility of pure bluffing—messages containing fairness ideals different from one's own.

Secondly, since our experiment is a one-shot game, players might be matched with partners possessing certain luck and effort, resulting in shares calculated by multiple fairness ideals that are not significantly different. In such cases, referring to one's ideal or a different ideal in a message becomes indifferent if the distribution outcomes are similar. Consequently, bluffing is possible only if the ideal results in a share akin to one's preferred allocation.

Next, we examine the reactions of subjects assigned the role of allocator in Stage 2 to the asking messages from others. Figure 4.7 illustrates bar graphs depicting the reactions of allocators by type and contents. In each panel, the leftmost bars represent the number of subjects who maintained their decision both before and after reading the message. The middle bars indicate those who increased their share for the other after reading the message, while the rightmost bars show the number of subjects who decreased their share to the other after reading the message.

Significantly, liberal egalitarians exhibited a higher percentage of subjects who retained their share without change compared to other types. On the other hand, egoists had a higher percentage of subjects who decreased their share compared to other types. This finding aligns with the results in Table 4.4, where liberal egalitarians are the only type seemingly not influenced by the empathy treatment. This suggests that they are the least susceptible among the four types to external factors, such as empathy or communication, when making distribution decisions.

Moreover, egoists demonstrated a higher proportion of subjects who decreased their share after reading the message, but a similar proportion increased their share after reading the message. When considered alongside the results in Table 4.4, this implies that egoists are individuals most influenced by external factors such as empathy and communication.

4.4 Summary and Discussion

This study examines how promoting state empathy, which is expressed in specific contexts or situations, affects the distribution decision-making of subjects with different fairness ideals. By introducing the A(e) condition of Andreoni and Rao (2011) to our experiments, we promote state empathy in Stage 1 of the empathy treatments by allowing both the allocator and the receiver sides to make decisions and write non-binding messages to their partners under role uncertainty. After Stage 1 of the empathy treatment, on average, the subjects decreased their own shares. However, by comparing by type, we find that the egoistic allocators (type EGO) reduce their shares the most, followed by the strict egalitarian and the libertarian allocators (type L and type LE), liberal egalitarian maintain their distribution regardless of empathy. In addition, in Stage 2 of the empathy treatment, the roles are assigned and the allocators make final distribution decisions after reading the receivers' asking message. After stage 2, we find no differences in the allocators' distribution decisions before and after the receiver's asking messages for any type. This result is contrary to those of Andreoni and Rao (2011); that is, our results show that the receiver's asking messages do not seem to have the power to change the allocator's distribution decision.

The effect of state empathy differs among subjects with different fairness ideals. In the case of type EGO, promoting state empathy significantly increases the altruism of distributions to others, a result observed in many previous studies. However, for type SE and type L subjects, the effect of state empathy is smaller than that observed for type EGO. Moreover, type LE subjects are not affected by state empathy. Instead, we find that for type SE and type L, state empathy tends to move the distribution closer to their ideal distributions according to their fairness ideals. In other words, state empathy reduces selfishness in general, allowing type EGO individuals to reduce their tendency to take all the endowment selfishly and move towards a more altruistic distribution. It also causes type SE and L individuals to reduce their selfishness and regulate themselves to distribute more in line with their fairness ideals.

While there was no significant overall effect for the messages, the egoist had the highest percentage of affected subjects among the four types. The subjects were likely to mention their own fairness ideals when writing the messages, but because they were cheap talk, we cannot rule out the possibility of bluffing in the messages that differed from the actual distribution decision.

Among the various types of empathy, state empathy refers to empathy promoted in a specific situation. However, we did not measure the trait empathy as an individual's ability to be empathetic to others. Therefore, there remains a possibility that subjects who are high in the trait empathy are classified as type LE and are not affected by state empathy. Moreover, as shown in Section 3.5.4, there could be a gender effect on the distribution of fairness ideal types and distribution decisions (Miller & Ubeda, 2012; Sharma, 2015). Therefore, the following two points remain for future work: (1) finding out whether the trait empathy affects individuals' fairness ideals and distribution decisions, and (2) analyzing the content of messages in the empathy treatment to determine whether asking messages containing specific content affects distribution decisions by fairness ideal

type.

In conclusion, we suggest that more analyses are needed on the positive effect of empathy on altruistic behaviors and distributional preferences. If the endowments to be distributed are determined by each individual's efforts and luck, state empathy may primarily promote altruism in certain types of people, such as type EGO individuals. In other words, altruistic behaviors may be induced when people who do not have a specific fairness ideal related to distribution and who aim to maximize their profit are affected by state empathy through a situation that allows them to put themselves in another's shoes.

Moreover, individuals who have their own ideals regarding fair distribution appear to prefer distributing according to these ideals regardless of whether state empathy is promoted. Therefore, empathetic campaigns for redistribution policies may be less effective for some people. The result of our type estimation shows that about 39.5% of the population is classified as egoists (type EGO), and they react to empathy-promoting settings. However, our results imply that campaigns intended to induce altruism may not work as effectively for other types of individuals, such as liberal egalitarians (type LE). These individuals may prioritize fair distribution according to their own fairness ideals over empathy for those who are at a disadvantage.

Recently, the problem of inequality has been highlighted as the cause of many social problems; moreover, the establishment and implementation of redistribution policies have emerged as important challenges in individual societies. With the goal of popularizing redistribution policies, campaigns that induce state empathy by showing the stories or situations of people in disadvantageous positions have become common. However, the results of this study show that some people believe that whether an initial distribution was done rightfully on a fairness principle is more important than redistribution policies or humanitarian aid projects that use empathetic campaigns. Furthermore, we suggest that such empathetic campaigns that provide opportunities to empathize with disadvantaged people may not always be effective, and that the effects of state empathy may be overestimated and overused. We also emphasize the fact that fair distribution is a more important issue for some people than an increase in altruism caused by empathetic promotion.

Chapter 5

Conclusion

5.1 Summary and Discussions

The distributional preferences explored in this thesis are of substantial importance in the field of behavioral economics, particularly in the context of increasing global inequality. Individuals' choices and behaviors are influenced not only by their own monetary payoffs but also by how their payoffs compare to others and the overall distribution within society. The main objective of this thesis was to investigate the heterogeneity in distributional preferences through a study of a public goods game with costly punishment and a dictator game involving real effort tasks.

In Chapter 2, the focus was on understanding the motivations underlying costly punishment in social dilemmas, specifically examining reciprocity and inequality aversion. To differentiate between these motivations, an element of uncertainty was introduced into the payoff function of a public goods game, preventing participants from accurately calculating others' contributions based on their payoffs. Additionally, a random income game was included to test the validity of the classification by removing intentionality from the contribution decisions. The findings revealed heterogeneity in punishment motivations, resulting in distinct types of punishers: self-interested, reciprocal, inequality-averse, and an "other" type characterized by inconsistent punishment patterns. Moreover, in the random income game, overall punishment tended to be lower, but only reciprocal punishment decreased with the presence of intention. Notably, the reciprocal type displayed strong punishment tendencies towards free-riding behaviors while also exhibiting some punishment for payoff inequality. These findings emphasized the critical role of inequality aversion as a primary motivation for punishment, with a substantial number of individuals relying solely on inequality aversion. Additionally, it was observed that individuals primarily motivated by reciprocity also demonstrated concern for payoff inequality, although reciprocity remained their main motivation.

Chapter 3 aimed to validate the findings presented by Cappelen et al. (2007) regarding the pluralism of fairness ideals. Their study, based on the dictator game with production, suggested the coexistence of three fairness ideals: strict egalitarianism, libertarianism, and liberal egalitarianism. However, considering the characteristics of the dictator game, where the egoistic decision of taking all endowments is rational, the presence of egoistic behavior cannot be ignored. In this study, modified models are employed to estimate the distributions based on different fairness ideals. We introduce a pure egoistic type who does have ideal for fairness to the model of Cappelen et al. (2007) and assume four types: strict egalitarians, libertarians, liberal egalitarians, and egoists. Additionally, we also suggest another modified model that assumes that individuals following different

fairness ideals assigned varying weights to fairness, and the weight parameter was separated according to the three fairness ideals. As a result, around 40% of subjects are classified as egoists, representing a substantial majority. However, notwithstanding this majority, the three fairness ideal types proposed by Cappelen et al. (2007) still coexist. Moreover, by comparing the values of the weight parameter across the fairness ideals, we find that strict egalitarians assign the highest weight to their ideal, followed by liberal egalitarians, with libertarians having the lowest weight among the three types.

In Chapter 4, the study examined how promoting state empathy, expressed in specific contexts or situations, influenced the decision-making process of individuals with different fairness ideals. The study initially categorized subjects into three fairness ideal types: egoists, libertarians, and liberal egalitarians, using the choice model of Cappelen et al. (2007) and the estimation method of Kwon and Funaki (2022). To promote state empathy, the A(e) condition of Andreoni and Rao (2011) was introduced, allowing both allocators and receivers to make decisions and exchange non-binding messages under role uncertainty. On average, subjects exhibited a reduction in their shares in response to the promotion of state empathy. However, when comparing the different types, it was observed that egoistic allocators, who lacked a fairness ideal, displayed the most significant reduction in their shares. In contrast, the impact of state empathy on other three types, such as strict egalitarians, libertarians, and liberal egalitarians, was relatively weaker compared to egoists. Instead, state empathy tended to align the distributions of strict egalitarians and libertarians closer to their ideal distributions based on their fairness ideals.

Now, I propose several implications derived from this thesis. Firstly, it emphasizes the importance of discussing distributional preferences in conjunction with heterogeneity. The findings from Chapter 2 highlight that distributional preferences are a crucial aspect of heterogeneous societies, coexisting alongside other social preferences like reciprocity. Therefore, it is important to analyze these behaviors from a heterogeneity perspective, even when observing the same pro-social behaviors, such as costly punishment without future benefit. Neglecting heterogeneity when examining costly punishment behavior, particularly in the presence of uncertainty, may lead to misunderstandings, such as misinterpreting punishments aimed at addressing inequality as antisocial punishments.

In the real world, good intentions or suitable processes do not always result in favorable outcomes, and the importance of outcome fairness can be overshadowed by the overestimation of process justice or the intentions behind certain behaviors. Thus, the results presented in Chapter 2 emphasize the significance of distributional justice by demonstrating that outcome fairness should be considered an important factor for certain individuals. These individuals may prioritize outcome fairness regardless of the presence or absence of cooperative behavior.

Secondly, the results from Chapter 3 highlight the importance of recognizing the presence of heterogeneity within distributional preferences when individual effort and luck are involved in jointly production. It is important to acknowledge that individuals advocating for a "fair share" may reveal different distributions according to their respective fairness ideals. The question of what defines a fair distribution lacks a universally definitive answer but rather requires reaching a consensus through persuasion and discussion. Thus, achieving social consensus among individuals with plural fairness ideals on what factors should be considered in determining a fair distribution becomes challenging without understanding and acknowledging the pluralism of fairness ideals.

Lastly, Chapter 4 suggests that empathy, a well-known factor for promoting altruistic giving, may be less effective for individuals who already perceive the distribution status as fair based on their ideal. This highlights the need to explore additional factors that

foster altruistic behavior. These factors could include the receiver's external environment, such as their income level and the social distance between the receiver and the allocator. Additionally, the allocator's education level and age may also play a role. Understanding how these factors relate to individuals' own fairness ideals and comparing the effect of each factor by fairness ideal type is important, as shown in our study. For example, one type may increase their share to others in response to state empathy, which is facilitated by being presented with the external environment of others, such as their income or standard of living. On the other hand, another type may be influenced by the allocator's social characteristics or solely adhere to their fairness ideal without being affected by external factors. By analyzing how these factors affect the distribution decisions of each type and understanding the characteristics of each type, I can foster a mutually respected consensus on fairness ideals within our society.

5.2 Future Works

Several topics within the study of distributional preferences require further investigation. Firstly, additional experimental designs are needed to verify the existence of strict egalitarian allocators. For instance, an experiment could be designed to examine whether individuals classified as strict egalitarians would display egoistic allocations even in three-player games. By estimating β^{SE} in three-player games, which represents the weight assigned to fairness for strict egalitarians (as seen in the choice model described in Section 3.3 and Section 3.4.2), I can determine if it corresponds to the magnitude of the egoistic choice that takes all the share. This verification would help reaffirm our claim that strict egalitarianism is practically nonexistent. Furthermore, it would be interesting to explore if strict egalitarian distributions would emerge when a third party decides in the same situation, where jointly production is determined by individual efforts and luck.

Secondly, our findings indicate gender differences in distributional preferences, highlighting the need to investigate the relationship between an individual's fairness ideal type and the gender characteristic of empathy sensitivity. Previous studies have shown that females tend to be more fairness-seeking than males (Kamas, Preston, & Baum, 2008; Sharma, 2015) and exhibit greater sensitivity to empathy (Christov-Moore et al., 2014; Croson & Gneezy, 2009; Kamas & Preston, 2021). However, our study revealed that females have a higher proportion of individuals with distributional preferences, particularly in terms of the liberal egalitarian type, compared to males, as illustrated in Figure 3.5. Additionally, Chapter 4 demonstrated that only egoists, constituting approximately half of the males, responded to state empathy (Figure 4.2 and Table 4.4). Therefore, it is important to investigate whether an individual's distribution changes based on the effect of empathy on their fairness ideal type or based on the sensitivity to empathy exhibited by females. This analysis will allow us to explore the interactions among gender, empathy, and our classification of fairness ideal types. Moreover, it may emerge that so-called gender stereotypes are invalidated by other social preferences, in this case, distributional preferences, which influence decision-making.

Appendix A

Experimental Instructions for Chapter 2

Thank you for your participation in this experiment.

You will participate in an experiment of an individual decision making. After the instructor reads this instruction, you will make decisions to make money. All the decisions should be done by inputting them into a computer in front of you. Your earnings depend on decisions by you and others. Your personal information, your decision and earnings are not known to the others. During the experiment, do not talk with others. If you have questions, let us know by raising your hand. Your mobile phone and pens should be in your bag.

There are two experiments EX1 and EX2 today. In EX1, there are 10 rounds which consist of 2 stages. You will be a member of a group of 4 persons. The members of the group are randomly chosen by a computer before starting each round. Then members of your group will be changed in each round. You do not know IDs of the other members of your group.

Next we explain your decision-making situation in each round of EX1.

EX1

1) Stage 1

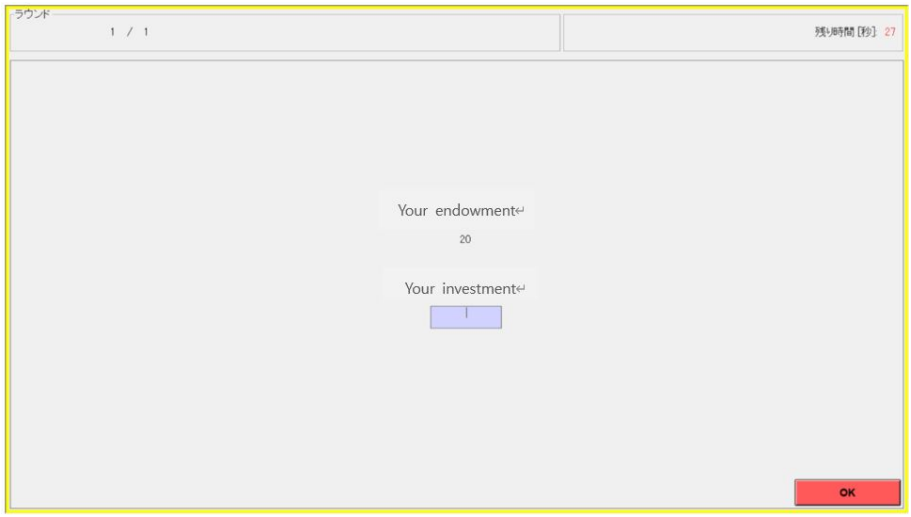


Figure A.1: Screenshot of Stage 1

Figure A.1 is a screenshot of stage 1. You will get 20 tokens at the beginning of each round. You will do some joint work with your group members. Then you choose number of tokens from your endowment 20 as an investment to your group's joint work. The rest of the amount of the endowment will be kept in your pocket. After all the members decide the input of an investment, your group's total amount of the investment will be doubled and divided equally among all the four members. Choose your number of tokes to invest and put the number into the input box. The number should be an integer between 0 and 20. Then click OK.

Additionally, when the payoff of stage 1 is calculated, all the group members get their own random increments. The range of increment is from -8 to +8. Here increment of -8 means decrement of 8. Your increment might be different from the other members.

As a result, your payoff of stage 1 follows this equation:
Payoff of stage 1 = 20 – your investment to joint work + (2×your group's total investment)/4 + increment (-8 to +8).

2) Stage 2

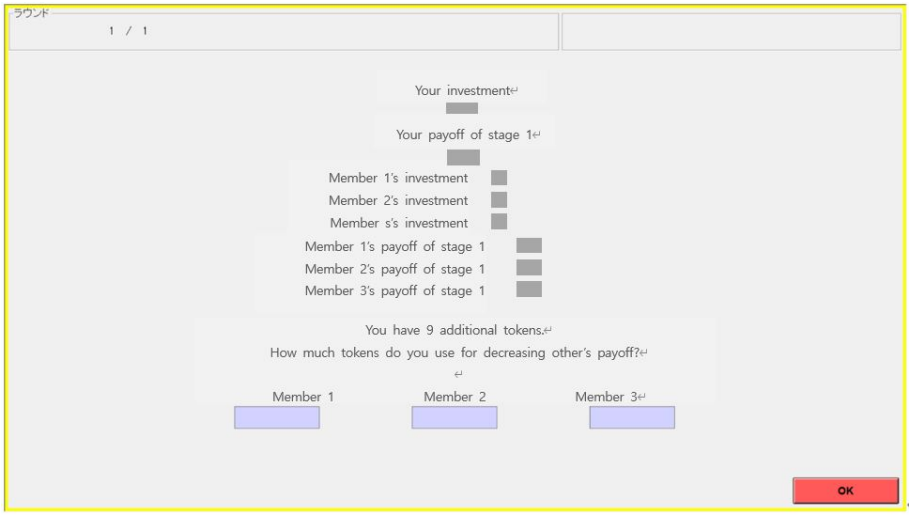


Figure A.2: Screenshot of Stage 2

Figure A.2 is a screenshot of stage 2. In this stage, you have 9 additional tokens. You can use these additional tokens to decrease other members' payoff. Each one token decreases 3 tokens of the target's payoff of stage 1. Additional tokens which you did not use are added to your pocket again.

Please input numbers of tokes between 0 and 3 into the input boxes of the other three members. Then click OK. You do not know who used token in this stage, and the other members do not know your input at this stage, too.

3) Result

This is a result screen. Your final payoff of this round follows this equation:
Final payoff of this round = payoff of stage 1 + 9 – used total tokens to other members - 3× total tokens from other members.

On the result screen you will find all the data After checking your result, click OK. When everyone has finished checking own result of this round, next round will start. After the end of 10 rounds, 3 rounds out of 10 will be chosen randomly and you will get cash by converting 10 Japanese yen per token based on your earing tokens. You will get this

Figure A.3: Screenshot of Result

amount of money plus 700 Japanese yen for a participation fee. You will get cash from EX2 later.

This is the end of an explanation of EX1. If you have questions, please let me know by raising your hand. One of the instructors will come to you. During the experiment you can read this instruction anytime. You can find a remaining time on the upper right corner of the screen. Please make your decision before the time limit.

Instruction of EX2. (This part of the instruction was distributed after EX1 was finished.)

In this experiment, each round consists of two stages and it repeats several times. Your group is reformed randomly at the beginning of each round.

1) Stage 1

You have 20 tokens as an endowment at the beginning. However, in this experiment, a computer chooses your investment and your group member's investments randomly. No one can make own investment decision. Decisions of stage 1 are automatically made and the payoff of this stage is given by the same equation as stage 1 of EX1. Then all the participants move on the next stage.

2) Stage 2

This stage is the same as stage 2 of EX1. You can use additional 9 tokens to decrease others' payoff of stage 1. Based on a computer's decision of stage 1, please input numbers of tokes between 0 and 3 into the input boxes of the other three members. Then click OK. You do not know who used token in this stage, and the other members do not know your input at this stage, too.

3) Result

You will find the final payoff of this round as EX1. Check your result, then click OK. After the end of all the rounds, 1 round will be chosen randomly and you will get cash by converting 10 Japanese yen per coin based on your earning tokens additionally. Your final cash amount is the sum of cash from EX1, EX2 and the participation fee.

Appendix B

Antisocial Punishment in Experiments of Chapter 2

To check the existence of antisocial punishment, we propose new models.

$$p_{ij} = \beta_1 + \beta_2 \min\{Con_j - Con_i, 0\} + \beta_3 \max\{Pay_j - Pay_i, 0\} + \beta_7 \max\{Con_j - Con_i, 0\} \quad \text{(M5)}$$

$$p_{ij} = \beta_4 + \beta_5 \min\{Con_j - GCon_{-j}, 0\} + \beta_6 \max\{Pay_j - GPay_{-j}, 0\} + \beta_8 \max\{Con_j - GCon_{-j}, 0\} \quad \text{(M6)}$$

The new variables, $\max\{Con_j - Con_i, 0\}$ and $\max\{Con_j - GCon_{-j}, 0\}$, imply antisocial punishments. As our type estimation, subjects who have significant β_7, β_8 or both carry out antisocial punishment. Because the antisocial punishment is defined as punishments for prosocial *behaviors* (Herrmann, Thoni, & Gachter, 2008; Thöni, 2014), we rule out punishments for those who earn lower payoffs than own.

There are ten subjects (7.4 % of all participants) who have significant β_7, β_8 or both. This percentage is very low compared to to the results in previous studies[1]. Table B.1 shows results of subjects who carry out both prosocial and antisocial punishment at the same time. Among Type R, four subjects performed significant antisocial punishment (10% of Type R), four among Type IA (11.4% of Type IA), and two among Type O (22.2% of Type O).

Moreover, in the Table B.1, the antisocial punishment variables ($\max\{Con_j - Con_i, 0\}$ and $\max\{Con_j - GCon_{-j}, 0\}$) show both a positive correlation (subjects 29, 58, 60, 67, 83, and 115) and a negative correlation (subjects 20, 32, 114, and 124) with the level of punishment. This result means that, unlike the prosocial punishment, people who carry out the antisocial punishment spend their endowment for punishment regardless of the difference in contributions between their own and the targets' who contribute more than them, and this is similar to the results of Herrmann et al. (2008). However, overall, the significance and the magnitude of antisocial punishment variables are lower than the prosocial punishment variables.

Therefore, in our study, since it is difficult to consider the antisocial punishment as the motivation for punishment, such as reciprocity and inequality aversion, we do not classify subjects who carry out antisocial punishments as a distinct type.

[1] It is 15% in Fehr and Gächter (2002) and 22% in Gächter and Herrmann (2009) as calculated by Thöni (2014).

Table B.1: Antisocial Punisher

	Type R				Type IA				Type O	
Dependent variable: Punishment level	Subject 20	Subject 29	Subject 67	Subject 114	Subject 60	Subject 83	Subject 115	Subject 124	Subject 32	Subject 58
M5 Individual comparison										
$\min\{Con_j - Con_i, 0\}$	-2.229^{***} (0.567)	-0.404 (0.328)	-2.966^{***} (1.005)	-2.471^{***} (0.721)	-0.112 (1.757)	-2.651^{**} (1.032)	-2.543^{*} (1.302)	(omitted)	0.711 (3.030)	-0.182 (0.496)
$\max\{Pay_j - Pay_i, 0\}$	-0.302 (0.800)	0.181 (0.279)	0.664 (0.803)	1.306 (0.949)	3.997^{***} (0.906)	2.597^{**} (1.070)	2.550^{***} (0.857)	2.140^{*} (1.213)	0.555 (2.045)	-0.061 (0.349)
$\max\{Con_j - Con_i, 0\}$	-0.356^{*} (0.200)	0.071 (0.225)	0.432 (0.382)	-5.079^{**} (2.139)	1.003^{**} (0.373)	2.997^{**} (1.115)	1.244^{**} (0.598)	-0.765^{*} (0.411)	-1.967 (1.498)	0.320^{**} (0.136)
constant	2.488^{***} (0.544)	0.374 (0.289)	2.723^{***} (0.966)	3.305^{***} (0.641)	0.026 (1.680)	3.394^{***} (0.873)	2.449^{*} (1.227)	0.545^{***} (0.160)	0.959 (2.945)	0.151 (0.485)
M6 Group comparison										
$\min\{Con_j - GCon_{-j}, 0\}$	-1.544^{***} (0.517)	-0.735^{***} (0.225)	-1.438^{*} (0.840)	-3.944^{***} (1.059)	1.519 (1.081)	-1.979^{*} (1.081)	-1.433^{*} (0.794)	-1.660 (1.146)	0.911 (2.065)	-0.079 (0.267)
$\max\{Pay_j - GPay_{-j}, 0\}$	-0.435 (0.578)	0.394^{*} (0.193)	1.334^{**} (0.524)	-0.015 (1.095)	2.464^{***} (0.759)	3.199^{***} (0.905)	2.748^{***} (0.599)	0.723 (0.829)	0.953 (1.780)	-0.008 (0.238)
$\max\{Con_j - GCon_{-j}, 0\}$	-0.273 (0.280)	0.251^{*} (0.138)	0.785^{*} (0.447)	-2.384^{**} (0.948)	0.669^{*} (0.384)	1.616^{*} (0.876)	1.124^{*} (0.546)	-0.477 (0.478)	-3.279^{**} (1.493)	0.317^{**} (0.143)
constant	1.707^{***} (0.491)	0.613^{***} (0.207)	1.073 (0.782)	5.117^{***} (1.010)	-1.404 (1.008)	2.785^{***} (0.970)	1.282^{*} (0.737)	1.988^{*} (1.112)	0.861 (1.956)	0.052 (0.252)

Note: All variables except the dependent variable are normalized. Standard errors are given in parentheses. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Appendix C

Punishment Behaviors in Experiments of Chapter 2

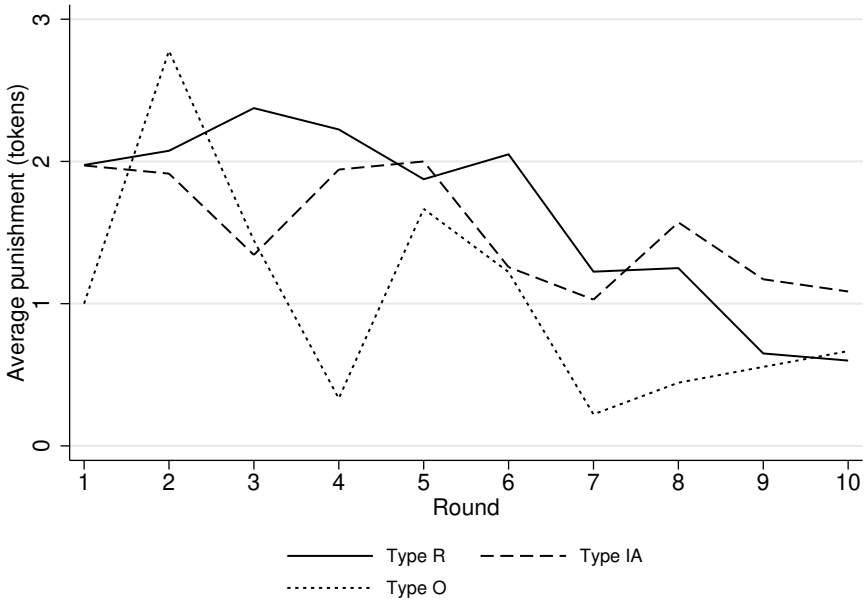


Figure C.1: Average Punishment across All Rounds

Figures C.1 and C.2 show trends in punishment behavior by round and by type. We observe that the level of punishment and the percentage of subjects who punish decreases as the round continues for all types. There are no significant differences between types in punishment trends.

To check the existence of counter punishment, we propose new models.

$$p_{ij}^t = \beta_1 + \beta_2 \min\{Con_j - Con_i, 0\} + \beta_3 \max\{Pay_j - Pay_i, 0\} + \beta_7 \sum_{j \neq i} p_{ji}^{t-1} \quad \text{(M7)}$$

$$p_{ij}^t = \beta_4 + \beta_5 \min\{Con_j - GCon_{-j}, 0\} + \beta_6 \max\{Pay_j - GPay_{-j}, 0\} + \beta_8 \sum_{j \neq i} p_{ji}^{t-1} \quad \text{(M8)}$$

The new variable, $\sum_{j \neq i} p_{ji}^{t-1}$, implies the total punishment received in the previous round. We normalize these variables, $\min\{Con_j - Con_i, 0\}$, $\min\{Con_j - GCon_{-j}, 0\}$, $\max\{Pay_j -$

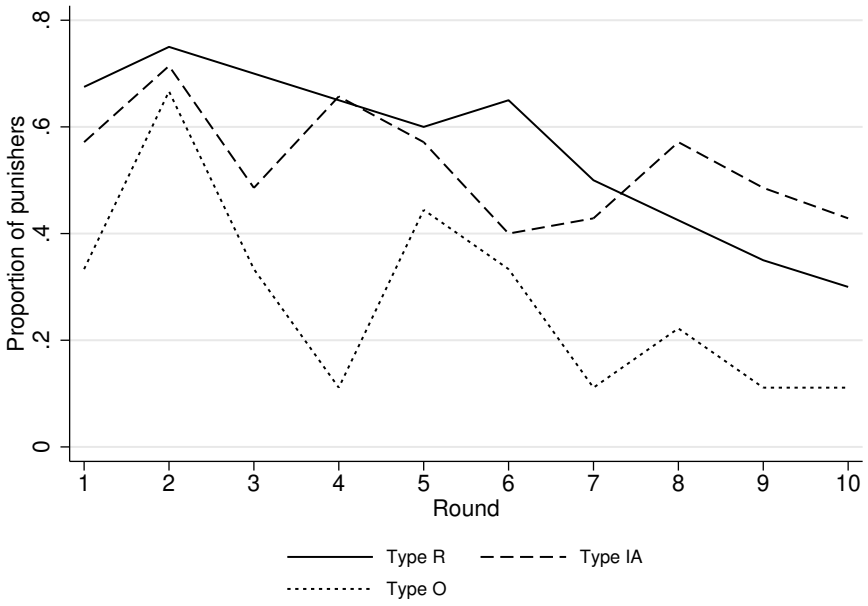


Figure C.2: Proportion of Punisher across All Rounds

Table C.1: Panel Tobit Regression Results for Punishment Behaviors.

Dependent variable: Punishment level	Type R	Type IA	Type O
M7 Individual comparison			
$\min\{Con_j - Con_i, 0\}$	-9.2136^{***} (0.6844)	-1.7735^{**} (0.6721)	0.4394 (2.9729)
$\max\{Pay_j - Pay_i, 0\}$	1.2171 (0.6401)	9.2019^{***} (0.8219)	2.7122 (3.6168)
$\sum_{j \neq i} p_{ji}^{t-1}$	0.2318^{**} (0.0781)	0.0281 (0.0722)	1.0244^{*} (0.4251)
Constant	5.0262^{***} (0.5852)	-1.2611 (0.7186)	-6.8908^{*} (3.3453)
M8 Group comparison			
$\min\{Con_j - GCon_{-j}, 0\}$	-9.2347^{***} (0.7180)	-1.1475 (0.6583)	0.9674 (3.3740)
$\max\{Pay_j - GPay_{-j}, 0\}$	1.7935^{**} (0.5964)	8.2906^{***} (0.6996)	5.7355 (3.1871)
$\sum_{j \neq i} p_{ji}^{t-1}$	0.0940 (0.0838)	-0.0234 (0.0705)	0.9660^{*} (0.4167)
Constant	5.5473^{***} (0.6281)	-1.7914^{*} (0.4065)	-7.7559^{*} (3.7770)
Number of observation	1080	945	243

Note: Standard errors are given in parentheses. We normalize all variables except variables for punishment levels. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

$Pay_i, 0\}$ and $\max\{Pay_j - GPay_{-j}, 0\}$, to have values between 0 and 1 in order to compare their magnitude of coefficients.

Table C.1 shows the panel tobit regression results for punishment behaviors in the

baseline game. Interestingly, Type R and Type O subjects increase their punishment if they were punished in the previous round. In our experiment, we have a stranger matching condition where the group composition changes every round, so there is a high probability that the person who punished him or her is not currently in his or her group. Nevertheless, Type R subjects tend to increase their revenge punishment slightly when comparing their own and others' contributions (Model M7), as shown by the results for variable $\sum_{j \neq i} p_{ji}^{t-1}$, which is significant at the 1% level but very small in magnitude. However, in Model M8, variable $\sum_{j \neq i} p_{ji}^{t-1}$ is not significant, suggesting that they do not consider revenge punishment when the group's contribution level is used as a criterion for punishment. In addition, Type O subjects tend to perform only revenge punishments that are not related to the other's contribution level or payoff, as indicated by the 5% level significance of variable $\sum_{j \neq i} p_{ji}^{t-1}$ in both Model M7 and Model M8.

This result shows additional characteristics of each type. First, Type R, who is motivated by reciprocity, punishes an unspecified target with a slight but significant punishment of revenge, suggesting that Type R may take the punishment received as malicious. Second, Type IA subjects - those who use punishment to reduce inequality in outcomes - do not seem to care about whether they have been punished by others, i.e., about the intentions of others toward them, but rather about the inequality itself or the cause of that inequality. Third, Type O, who exhibits inconsistent punishment behavior, motivated neither by the other's behavior nor by payoff inequality, is most likely motivated by retaliatory punishment among other social preferences - spite, envy, competitiveness, etc.

Appendix D

Experimental Instructions for Chapters 3 and 4

Thank you for your participation in the experiment today.

You will participate in an experiment of individual decision-making. After the instructor reads this instruction, you will make decisions to make money. All the decisions are made by inputting them into a computer in front of you. During the experiment, do not talk with others. If you have questions, let us know by raising your hand. Your mobile phone and pens should be in your bag.

Your payment is determined by the decision of you and other participants in addition to the participation fee of 800 yen. Your personal information, your decisions, and your earnings are not known to others.

This experiment consists of two parts. First of all, Part 1 consists of two stages. After finishing Stage 1, you will be given an experiment manual for Stage 2.

Stage 1

Figure D.1: Screenshot of Stage 1

You will participate in one task. Figure D.1 is a screenshot of the task. In this figure, there are 48 sliders. Your work is to align the bar on 0, the left side of each slider, with 50, the center of the slider.

In other words, your goal is to move the bar from 0 to 50 more accurately within a limited time. Score 1 point for each slider that aligns with 50. First of all, you will participate in the practice task with a time limit of 120 seconds. The results of this practice task are not related to the payment. The remaining time is displayed on the screen in front of the classroom. When there are 30 seconds left, and when there are 10 seconds left, we will inform you of the remaining time verbally. Once the practice tasks are finished, the production tasks used for decision-making of Stage 2 begin. The time limit is 150 seconds.

Your output is determined by the product of the rate of return and the points you get from the task. Your rate of return, one of 2 or 4, will be randomly assigned to you after the production task is finished.

$$\text{Your output} = \text{your rate of return} \times \text{your points}.$$

Check your rate of return and your output from the result screen of Stage 1.

Stage 2

The sum of the output of Stage 1 earned by you and your partner will be the product of your joint venture. In Stage 2, you will make a decision in a pair. Your payment is determined by the results of Stage 2. This stage repeats several times, each round pairing you with a new partner. You will never be paired with the same partner. After Stage 2, two rounds out of all the decision-making rounds are chosen to be the payment of this experiment.

Figure D.2: Screenshot of Stage 2

In this Stage 2, you will participate decision-making to distribute the sum of the output of Stage 1 earned by you and your partner (the group's total product) between you and your partner. Figure D.2 is a screenshot of Stage 2.

On the upper left side of the screen, 'your points from the task', 'your rate of return', and 'your output' are displayed. On the upper right side, 'your partner's points from the task', 'your partner's rate of return', and 'your partner's output' are displayed. At the bottom of the screen, you will see your group's total product.

Please distribute 'your share' and 'your partner's share' from your group's total production of this round and input them in integers. Please input each number so that the

sum of the two numbers is your group's total product. When your distribution decision is completed, please click the OK.

This decision-making will be repeated several times. After Stage 2, two rounds are randomly chosen, and one of your decisions or your partner's decisions in those rounds is randomly chosen. The sum of your points from these two decisions will be converted into cash at the rate of 1 point = 3 yen, and your final payment is the sum of cash from the experiment and the participation fee.

(Instruction for the baseline treatment of Part 2)

In Part 2, we will use the results of Stage 1 of Part 1. This part consists of one stage. You will be randomly paired with another participant by the computer. Figures D.3 and D.4 are screenshots of Part 2.

Figure D.3: Screenshot of Allocator

Figure D.4: Screenshot of Receiver

In Part 2, you and your partner will share the roles of the allocator and the receiver. The role will be chosen randomly by the computer.

Figure D.3 is a screenshot of the information displayed to the allocator, and Figure D.4 is a screenshot of that displayed to the receiver. On the upper-left side of the screen, you will see "Your points from the task", "Your rate of return" and "Your output" following your result of Stage 1 of Experiment 1. On the upper-right side, "Your partner's points from the task", "Your partner's rate of return", and "Your partner's output" will be displayed. At the bottom of the screen, you will see your group's total product.

If you are the allocator (see Figure D.3), you will own your group's total product and be able to freely distribute this product between you and your partner. Please input your share and your partner's share as integers. Please input the sum of the two numbers as the total product.

If you are the receiver (see Figure D.4), you will not own anything and follow your partner's decision. You will not be able to make any decisions. Please predict how many points your partner will share with you and input the number in the box below. After inputting your prediction, click the OK button and wait until your partner's decision is complete. Your prediction will not affect your partner's distribution decision.

Your points from Part 2 will be converted into cash at the rate of 1 point = 3 yen, and this will be your payment for Part 2. The sum of the payments for Part 1 and Part 2 will be your total payment for this experiment.

(Instruction for the empathy treatment of Part 2)

In Part 2, we will use the results of Stage 1 of Part 1. This part consists of two stages. You will be randomly paired with another participant by the computer.

Stage 1

Figure D.5: Screenshot of Stage 1

In Part 2, you and your partner will share the roles of the allocator and the receiver. The roles will be decided after Stage 1. Please make decisions considering the possibility of being either the allocator or the receiver in Stage 2.

At the beginning of Stage 1, you will not know whether you will be the allocator or the receiver. Please input your decision in each case as instructed on the left side for the allocator and on the right side for the receiver. Note that you have to decide for both cases.

On the left side of the top of the screen, you will see “Your points from the task”, “Your rate of return” and “Your output” following the result of Stage 1 of Experiment 1. On the right side, “Your partner’s points from the task”, “Your partner’s rate of return”, and “Your partner’s output” will be displayed.

If you are the allocator (on the left side of the screen), you will own your group’s total product and be able to freely distribute this product between you and your partner. “Your group’s total product” will be displayed in the upper portion of the screen in the left box. Please input your share and your partner’s share as integers in the small blue boxes. Please input the sum of the two numbers as the total product. In addition, it will be possible to send an explanation message about your decision to your partner at the same time. Please write a message of 150 characters or less to send to your partner in the large blue box at the bottom. If you write a message, please do not enter any information that identifies you, such as a seat number.

If you are the receiver (on the right side of the screen), you will not own anything and will follow your partner’s decision. However, it will be possible to send a request and a message. This message may influence your partner’s decision-making in the next stage if he or she is given the role of the allocator. “Your group’s total product” will be displayed in the upper portion of the screen in the box on the right. Please input your request share in integers in the small blue box and write a message in the large blue box at the bottom using 150 characters or less. If you write a message, please do not enter any information that identifies you, such as a seat number.

After inputting information on both the left and right sides, click the OK button.

Stage 2

Figures D.6 and D.7 are screenshots of Stage 2. After Stage 1 is finished, the roles are determined. If you are given the role of the allocator, Figure D.6 will be displayed, and if you are given the role of the receiver, Figure D.7 will be displayed. Make sure your role is shown at the top of the screen.

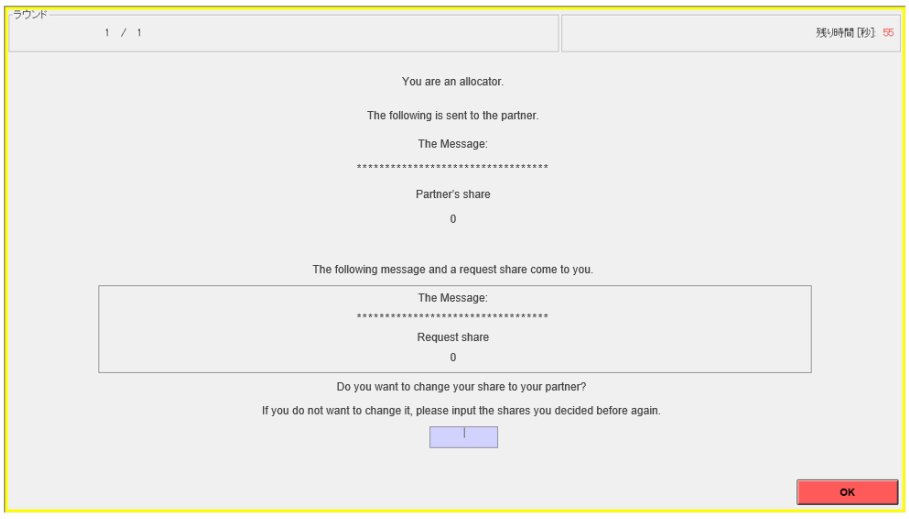


Figure D.6: Screenshot of Allocator

If you are given the role of allocator (Figure D.6), the message and partner share you chose as an allocator in Stage 1 will be shown on the upper part of the screen. Moreover, the request and message of the receiver will be displayed on the lower part of the screen. In this stage, the allocator will be able to change the distribution to the partner that was

Figure D.7: Screenshot of Receiver

determined in Stage 1. If you do not want to change it, please reinput the partner share you input in Stage 1. If you want to change it, input the chosen share to the partner in integers. Please click the OK button when the decision is complete.

If you are given the role of receiver (Figure D.7), the allocator's decision regarding your share and an explanation message will be shown on the upper part of the screen. Moreover, the request and message you sent to your partner will be displayed on the lower part of the screen. You will own nothing and follow your partner's decisions. You cannot make any decisions. Please predict how many points your partner will share with you and input the number in the box below. After inputting your prediction, click the OK button and wait until your partner's decision is complete. Your prediction will not affect your partner's distribution decision.

Your points from Part 2 will be converted into cash at the rate of 1 point = 3 yen, and this will be your payment for Part 2. The sum of the payments from Parts 1 and 2 will be your total payment for this experiment.

References

Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E. O., & Tungodden, B. (2017). Fairness and Family Background. *Politics, Philosophy & Economics*, *16*(2), 117–131.

Almås, I., Cappelen, A. W., Sørensen, E. O., & Tungodden, B. (2010). Fairness and the Development of Inequality Acceptance. *Science*, *328*(5982), 1176–1178.

Andreoni, J., Harbaugh, W., & Vesterlund, L. (2003). The Carrot or the Stick: Rewards, Punishments, and Cooperation. *American Economic Review*, *93*(3), 893–902.

Andreoni, J., & Rao, J. M. (2011). The Power of Asking: How Communication Affects Selfishness, Empathy, and Altruism. *Journal of Public Economics*, *95*(7-8), 513–520.

Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving. *Journal of Political Economy*, *125*(3), 625–653.

Andreoni, J., & Vesterlund, L. (2001). Which Is the Fair Sex? Gender Differences in Altruism. *The Quarterly Journal of Economics*, *116*(1), 293–312.

Ashraf, N., Camerer, C. F., & Loewenstein, G. (2005). Adam Smith, Behavioral Economist. *Journal of Economic Perspectives*, *19*(3), 131–145.

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, Punishment, and Cooperation: a Meta-Analysis. *Psychological Bulletin*, *137*(4), 594–615.

Basil, D. Z., Ridgway, N. M., & Basil, M. D. (2008). Guilt and Giving: A Process Model of Empathy and Efficacy. *Psychology & Marketing*, *25*(1), 1–23.

Batson, C. D. (2009). These Things Called Empathy: Eight Related but Distinct Phenomena. In J. Decety & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 3–15). Cambridge: The MIT Press.

Batson, C. D., Batson, J. G., Slingsby, J. K., Harrell, K. L., Peekna, H. M., & Todd, R. M. (1991). Empathic Joy and the Empathy-Altruism Hypothesis. *Journal of Personality and Social Psychology*, *61*(3), 413–426.

Batson, C. D., Coke, J. S., et al. (1981). Empathy: A Source of Altruistic Motivation for Helping. *Altruism and Helping Behavior: Social, Personality, and Developmental Perspectives*, 167–187.

Batson, C. D., Sager, K., Garst, E., Kang, M., Rubchinsky, K., & Dawson, K. (1997). Is Empathy-Induced Helping Due to Self--Other Merging? *Journal of Personality and Social Psychology*, *73*(3), 495–509.

Bault, N., Coricelli, G., & Rustichini, A. (2008). Interdependent Utilities: How Social Ranking Affects Choice Behavior. *PloS One*, *3*(10), e3477.

Becker, A. (2013). Accountability and the Fairness Bias: the Effects of Effort Vs. Luck. *Social Choice and Welfare*, *41*(3), 685–699.

Benenson, J. F., Pascoe, J., & Radmore, N. (2007). Children's Altruistic Behavior in the Dictator Game. *Evolution and human Behavior*, *28*(3), 168–175.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, *90*(1), 166–193.

Brañas-Garza, P. (2006). Poverty in Dictator Games: Awakening Solidarity. *Journal of Economic Behavior & Organization*, *60*(3), 306–320.

Bruttel, L., & Stolley, F. (2018). Gender Differences in the Response to Decision Power and Responsibility—framing Effects in a Dictator Game. *Games*, *9*(2), 1–16.

Cappelen, A. W., Hole, A. D., Sørensen, E. O., & Tungodden, B. (2007). The Pluralism of Fairness Ideals: An Experimental Approach. *American Economic Review*, *97*(3), 818–827.

Cappelen, A. W., Hole, A. D., Sørensen, E. O., & Tungodden, B. (2011). The Importance of Moral Reflection and Self-Reported Data in a Dictator Game with Production. *Social Choice and Welfare*, *36*(1), 105–120.

Cappelen, A. W., Moene, K. O., Sørensen, E. O., & Tungodden, B. (2013). Needs Versus Entitlements—an International Fairness Experiment. *Journal of the European Economic Association*, *11*(3), 574–598.

Cappelen, A. W., Sørensen, E. O., & Tungodden, B. (2010). Responsibility for What? Fairness and Individual Responsibility. *European Economic Review*, *54*(3), 429–441.

Carpenter, J. P., & Matthews, P. H. (2012). Norm Enforcement: Anger, Indignation, or Reciprocity? *Journal of the European Economic Association*, *10*(3), 555–572.

Casal, S., Güth, W., Jia, M., & Ploner, M. (2012). Would You Mind If I Get More? An Experimental Study of the Envy Game. *Journal of Economic Behavior & Organization*, *84*(3), 857–865.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental Methods: Measuring Effort in Economics Experiments. *Journal of Economic Behavior & Organization*, *149*, 74–87.

Charness, G., & Grosskopf, B. (2001). Relative Payoffs and Happiness: an Experimental Study. *Journal of Economic Behavior & Organization*, *45*(3), 301–328.

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.

Chen, Y., Zhu, L., & Chen, Z. (2013). Family Income Affects Children's Altruistic Behavior in the Dictator Game. *PloS one*, *8*(11), e80419.

Christian, R. C., & Alm, J. (2014). Empathy, Sympathy, and Tax Compliance. *Journal of Economic Psychology*, *40*, 62–82.

Christov-Moore, L., Simpson, E. A., Coudé, G., Grigaityte, K., Iacoboni, M., & Ferrari, P. F. (2014). Empathy: Gender Effects in Brain and Behavior. *Neuroscience & biobehavioral reviews*, *46*, 604–627.

Chuan, A., & Samek, A. S. (2014). "Feel the Warmth" Glow: A Field Experiment on Manipulating the Act of Giving. *Journal of Economic Behavior & Organization*, *108*, 198–211.

Cox, J. C., Friedman, D., & Gjerstad, S. (2007). A Tractable Model of Reciprocity and Fairness. *Games and Economic Behavior*, *59*(1), 17–45.

Croson, R., & Gneezy, U. (2009). Gender Differences in Preferences. *Journal of Economic literature*, *47*(2), 448–474.

Cuff, B. M., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A Review of the Concept. *Emotion Review*, *8*(2), 144–153.

Davis, M. H. (1983). Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian Motives in Humans. *Nature*, *446*(7137), 794–796.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment. *Economic Theory*, *33*(1), 145–167.

Dickinson, D. L., & Tiefenthaler, J. (2002). What Is Fair? Experimental Evidence. *Southern Economic Journal*, *69*(2), 414–428.

Drange Hole, A. (2011). Communication and Fair Distribution. *Rationality and Society*, *23*(2), 234–264.

Dreber, A., Ellingsen, T., Johannesson, M., & Rand, D. G. (2013). Do People Care About Social Context? Framing Effects in Dictator Games. *Experimental Economics*, *16*, 349–371.

Dufwenberg, M., & Kirchsteiger, G. (2004). A Theory of Sequential Reciprocity. *Games and Economic Behavior*, *47*(2), 268–298.

Dufwenberg, M., & Muren, A. (2006). Gender Composition in Teams. *Journal of Economic Behavior & Organization*, *61*(1), 50–54.

Edele, A., Dziobek, I., & Keller, M. (2013). Explaining Altruistic Sharing in the Dictator Game: The Role of Affective Empathy, Cognitive Empathy, and Justice Sensitivity. *Learning and Individual Differences*, *24*, 96–102.

Egas, M., & Riedl, A. (2008). The Economics of Altruistic Punishment and the Maintenance of Cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1637), 871–878.

Engel, C. (2011). Dictator Games: A Meta Study. *Experimental Economics*, *14*(4), 583–610.

Engelmann, D., & Strobel, M. (2004). Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments. *American Economic Review*, *94*(4), 857–869.

Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative Earnings and Giving in a Real-Effort Experiment. *American Economic Review*, *101*(7), 3330–3348.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving Forces Behind Informal Sanctions. *Econometrica*, *73*(6), 2017–2030.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing Theories of Fairness—Intentions Matter. *Games and Economic Behavior*, *62*(1), 287–303.

Falk, A., & Fischbacher, U. (2006). A Theory of Reciprocity. *Games and Economic Behavior*, *54*(2), 293–315.

Farrell, J., & Rabin, M. (1996). Cheap Talk. *Journal of Economic Perspectives*, *10*(3), 103–118.

Fehr, E., & Gächter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, *14*(3), 159–181.

Fehr, E., & Gächter, S. (2002). Altruistic Punishment in Humans. *Nature*, *415*(6868), 137–140.

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics*, *10*(2), 171–178.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic behavior*, *6*(3), 347–369.

Frohlich, N., Oppenheimer, J., & Kurki, A. (2004). Modeling Other-Regarding Preferences and an Experimental Test. *Public Choice*, *119*, 91–117.

Gächter, S., & Herrmann, B. (2009). Reciprocity, Culture and Human Cooperation: Previous Insights and a New Cross-Cultural Experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 791–806.

Gächter, S., Herrmann, B., & Thöni, C. (2004). Trust, Voluntary Cooperation, and Socio-Economic Background: Survey and Experimental Evidence. *Journal of Economic Behavior & Organization*, *55*(4), 505–531.

Gill, D., & Prowse, V. (2012). A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review*, *102*(1), 469–503.

Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York: Knopf Doubleday Publishing Group.

Heinz, M., Juranek, S., & Rau, H. A. (2012). Do Women Behave More Reciprocally Than Men? Gender Differences in Real Effort Dictator Games. *Journal of Economic Behavior & Organization*, *83*(1), 105–110.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *American Economic Review*, *91*(2), 73–78.

Herne, K., Hietanen, J. K., Lappalainen, O., & Palosaari, E. (2022). The Influence of Role Awareness, Empathy Induction and Trait Empathy on Dictator Game Giving. *PloS One*, *17*(3), e0262196.

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial Punishment Across Societies. *Science*, *319*(5868), 1362–1367.

Hilbe, C., & Traulsen, A. (2012). Emergence of Responsible Sanctions Without Second Order Free Riders, Antisocial Punishment or Spite. *Scientific Reports*, *2*(1), 1–4.

Iriberri, N., & Rey-Biel, P. (2011). The Role of Role Uncertainty in Modified Dictator Games. *Experimental Economics*, *14*(2), 160–180.

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The Role of Egalitarian Motives in Altruistic Punishment. *Economics Letters*, *102*(3), 192–194.

Kamas, L., & Preston, A. (2021). Empathy, Gender, and Prosocial Behavior. *Journal of Behavioral and Experimental Economics*, *92*, Article 101654.

Kamas, L., Preston, A., & Baum, S. (2008). Altruism in Individual and Joint-Giving Decisions: What's Gender Got to Do with It? *Feminist Economics*, *14*(3), 23–50.

Kijima, N. (1996). Cloninger's Seven Factor Model of Personality and Japanese Version of Temperament and Character Inventory. *Arch. Psychiatr. Diag. Clin. Eval.*, *7*, 379–399.

Kirchsteiger, G. (1994). The Role of Envy in Ultimatum Games. *Journal of Economic Behavior & Organization*, *25*(3), 373–389.

Kirman, A., & Teschl, M. (2010). Selfish or Selfless? The Role of Empathy in Economics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1538), 303–317.

Klimecki, O. M., Mayer, S. V., Jusyte, A., Scheeff, J., & Schönenberg, M. (2016). Empathy Promotes Altruistic Behavior in Economic Interactions. *Scientific reports*, *6*(1), 1–5.

Konow, J. (2000). Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions. *American Economic Review*, *90*(4), 1072–1091.

Krupka, E. L., & Weber, R. A. (2013). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of the European Economic Association*, *11*(3), 495–524.

Kwon, H., & Funaki, Y. (2022). Do Strict Egalitarians Really Exist? *WINPEC Working Paper Series*, *E2206*, 1–27.

Kwon, H., & Funaki, Y. (2023a). Heterogeneity in the Motivations for Punishment: An Experimental Study. *Available at SSRN 4441815*. Retrieved from `https://ssrn.com/abstract=4441815`

Kwon, H., & Funaki, Y. (2023b). The Empathetic Egoist: The Effect of Empathy on Plural Fairness Ideals. *Available at SSRN 4441637*. Retrieved from `https://ssrn.com/abstract=4441637`

Leibbrandt, A., & López-Pérez, R. (2012). An Exploration of Third and Second Party Punishment in Ten Simple Games. *Journal of Economic Behavior & Organization*, *84*(3), 753–766.

Levine, D. K. (1998). Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics*, *1*(3), 593–622.

List, J. A. (2007). On the Interpretation of Giving in Dictator Games. *Journal of Political Economy*, *115*(3), 482–493.

Masclet, D., & Villeval, M.-C. (2008). Punishment, Inequality, and Welfare: a Public Good Experiment. *Social Choice and Welfare*, *31*(3), 475–502.

Miller, L., & Ubeda, P. (2012). Are Women More Sensitive to the Decision-Making Context? *Journal of Economic Behavior & Organization*, *83*(1), 98–104.

Mittone, L., & Ploner, M. (2012). Asset Legitimacy and Distributive Justice in the Dictator Game: An Experimental Analysis. *Journal of Behavioral Decision Making*, *25*(2), 135–142.

Moffatt, P. G. (2015). *Experimetrics: Econometrics for Experimental Economics*. London: Palgrave Macmillan.

Nikiforakis, N., & Normann, H.-T. (2008). A Comparative Statics Analysis of Punishment in Public-Good Experiments. *Experimental Economics*, *11*(4), 358–369.

Nozick, R. (1974). *Anarchy, State, and Utopia* (Vol. 5038). New York: Basic Books.

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its Ultimate and Proximate Bases. *Behavioral and Brain Sciences*, *25*(1), 1–20.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review*, *83*(5), 1281–1302.

Raihani, N. J., & McAuliffe, K. (2012). Human Punishment Is Motivated by Inequity Aversion, Not a Desire for Reciprocity. *Biology Letters*, *8*(5), 802–804.

Revelt, D., & Train, K. (1998). Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level. *Review of Economics and Statistics*, *80*(4), 647–657.

Rifkin, J. (2009). *The Empathic Civilization: The Race to Global Consciousness in a World in Crisis*. New York: Penguin Group.

Rockenbach, B., & Milinski, M. (2006). The Efficient Interaction of Indirect Reciprocity and Costly Punishment. *Nature*, *444*(7120), 718–724.

Rodriguez-Lara, I., & Moreno-Garrido, L. (2012). Self-Interest and Fairness: Self-Serving Choices of Justice Principles. *Experimental Economics*, *15*(1), 158–175.

Ruffle, B. J. (1998). More Is Better, but Fair Is Fair: Tipping in Dictator and Ultimatum Games. *Games and Economic Behavior*, *23*(2), 247–265.

Selten, R., & Ockenfels, A. (1998). An Experimental Solidarity Game. *Journal of Economic Behavior & Organization*, *34*(4), 517–539.

Sharma, S. (2015). Gender and Distributional Preferences: Experimental Evidence from India. *Journal of Economic Psychology*, *50*, 113–123.

Singer, T. (2006). The Neuronal Basis of Empathy and Fairness. In *Empathy and fairness: Novartis foundation symposium 278* (pp. 20–40).

Singer, T., & Lamm, C. (2009). The Social Neuroscience of Empathy. *Annals of the New York Academy of Sciences*, *1156*(1), 81–96.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic Neural Responses Are Modulated by the Perceived Fairness of Others. *Nature*, *439*(7075), 466–469.

Smith, A. (2010). *The Theory of Moral Sentiments*. London: Penguin Classics. (Originally published in 1759)

Stata Corporation. (2005). *Stata Base Reference Manual: Release 9* (Vol. 3). Texas: Stata Corporation.

Thöni, C. (2014). Inequality Aversion and Antisocial Punishment. *Theory and Decision*, *76*(4), 529–545.

Ubeda, P. (2014). The Consistency of Fairness Rules: An Experimental Study. *Journal of Economic Psychology*, *41*, 88–100.

Urbanska, K., McKeown, S., & Taylor, L. K. (2019). From Injustice to Action: The Role of Empathy and Perceived Fairness to Address Inequality Via Victim Compensation. *Journal of Experimental Social Psychology*, *82*, 129–140.

Wang, A., Zhu, L., Lyu, D., Cai, D., Ma, Q., & Jin, J. (2022). You Are Excusable! Neural Correlates of Economic Neediness on Empathic Concern and Fairness Perception. *Cognitive, Affective, & Behavioral Neuroscience*, *22*(1), 99–111.

Xiao, E., & Bicchieri, C. (2010). When Equality Trumps Reciprocity. *Journal of Economic Psychology*, *31*(3), 456–470.

Zaki, J., & Ochsner, K. N. (2012). The Neuroscience of Empathy: Progress, Pitfalls and Promise. *Nature Neuroscience*, *15*(5), 675–680.

Zelmer, J. (2003). Linear Public Goods Experiments: A Meta-Analysis. *Experimental Economics*, *6*(3), 299–310.

Zizzo, D. J. (2003). Money Burning and Rank Egalitarianism with Random Dictators. *Economics Letters*, *81*(2), 263–266.

Zizzo, D. J., & Oswald, A. J. (2001). Are People Willing to Pay to Reduce Others' Incomes? *Annales d'Economie et de Statistique*(63/64), 39–65.