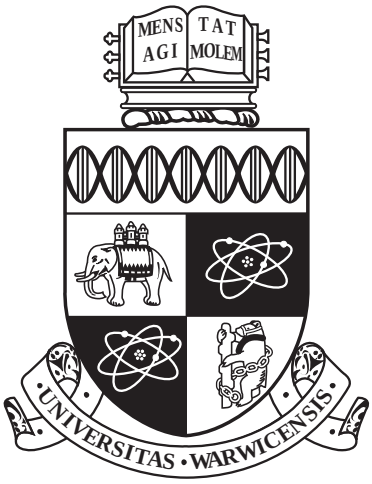


Essays on the Determinants and Effects of Social Preferences

by

Andrew John Siu

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Economics

September 2015

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my supervisors, Andrew Oswald and Robert Akerlof, for their guidance and support during my PhD research. I am particularly grateful to Chengwei Liu for his encouragement and help in applying for a research grant and collecting experimental data. For providing comments on early drafts of my chapters, I thank Dan Bernhardt, Niall Hughes, Chengwei Liu, Gordon Menzies, Aaron Nicholas, Eugenio Proto and Daniel Sgroi. I also greatly appreciate all the staff and research students in the Department of Economics who have provided a stimulating and positive environment for research.

I would also like to thank all the seminar participants at the University of Warwick and at the University of Stirling. For providing support in running the experiments, I thank the laboratory manager Alex Mushore at Warwick Business School and other helpers. Research grant from Behavioural Science Global Research Priorities is gratefully acknowledged. Warwick Economics Postgraduate Research Fellowship and other financial support from the Department of Economics are greatly appreciated.

I am eternally grateful to my mom and dad for their love and care, to my brother Daniel for his encouragement and advice, and to all my brothers and sisters in Christ for their prayers and shepherding. Also, I am grateful beyond words to my wife Cristina for all that she is and has done for me, including giving birth to a healthy and happy baby at the most difficult time of my PhD. Finally, I thank my dear Lord Jesus Christ for giving me the daily strength and wisdom I need to complete this thesis.

Declarations

I declare that any material contained in the thesis is my own work and has not been submitted for a degree at another university.

Andrew Siu

September 2015

Abstract

This thesis is designed as a contribution to the economics of social interaction with a focus on human emotions and thinking processes. The first two chapters are empirical and the third chapter is theoretical.

Chapter one examines the extent to which punishments are motivated by the emotion of anger or 'fairness' considerations. A laboratory experiment uses a multi-round game where the punisher could not be sure whether a selfish action of the punished may be 'excused' or not. The results show that subjects tend to inflict a harsher punishment as the proportion of observed selfish actions in previous rounds increases, after controlling for the current action. The data can further test competing hypotheses of two theories: norm compliance and spitefulness. One third of subjects punish an action that fails to comply with the norm, but none habitually punish a spiteful person regardless of the current action.

Chapter two investigates whether the individual tendency to think intuitively or deliberately can lead to altruistic giving or punishment. An online experiment uses a 40-item self-report questionnaire to measure individual reliance on intuitive feelings (Faith in Intuition) and personal tendency to engage in deliberate thinking (Need for Cognition). The results show that people who tend to think more deliberately are less prone to punish. An increase in the cost of punishing reduces both punishment and giving. High reliance on intuition is associated with greater sensitivity of punishment to a cost increase than to a cost decrease, which might be explained by loss aversion.

Chapter three develops a model of interdependent preferences in the presence of asymmetric information. The model explores the welfare consequences of permitting divorce. Suppose each player has a private value of the marriage and may or may not care about the partner's value. When divorce is possible, any player can use the threat of divorce to make demands on the other player, but it might also reveal one's own value of the marriage. A well-known theoretical result is that asymmetric information routinely leads to inefficient bargaining and divorce, but this model further shows that incorporating interdependent preferences can eliminate such inefficiencies. Thus, asymmetric information is not a sufficient condition for inefficient divorce; the lack of care about the partner is also necessary.

Chapter 1

Escalating Anger and Punishments: Testing Theories of Cumulative Anger

1.1 Introduction

Recent research suggests that, consistent with casual observation, anger can motivate people to punish others (Fehr & Gächter, 2002; Hopfensitz & Reuben, 2009). However, little has been done to study what factors can predict anger. Most studies focus on one-shot interaction, such as the ultimatum game, where anger is potentially triggered as an immediate reaction to an unsatisfactory proposal. By contrast, this study investigates to what extent a transgressor's history of norm violations can predict a punisher's present and future anger and punishment.

There is a common perception that history should matter: people tend to punish repeat violators more harshly than first-timers. In most countries, the law imposes harsher punishments for repeat offenders, and such legislation receives widespread public support (see Roberts, 1997). One explanation for this comes from a principal-agent model (Polinsky & Rubinfeld, 1991), in which the principal imposes a higher fine for the second offense in order to optimally deter repeat offenders. These increasingly severe and thus 'graduated' punishments that have a goal of optimizing deterrence can, in theory, be implemented without the emotion of anger, but anecdotal evidence suggests that escalating anger plays an important role in motivating escalating

punishments. For instance, an article by Trounstine (2014) reveals that the U.S. public's angry responses to brutal crimes have led to state and federal legislation. Notably, these horrific crimes trigger not only anger but also the occasional practice of naming a law after dead victims.[1]

This study examines the determinants of anger and punishments and how these observed patterns relate to existing theories. We conduct a laboratory experiment in which two subjects are matched and then play a multi-round game. In each round, the first player decides whether to give an equal share of a pot of money to the second player, and then the second player can decide how much to punish the first player. Asymmetric information is created between the two parties based on Durham's (1987) observation that 'repetitive criminal involvement indicates the existence of "hidden" attributes possessed by the offender.' In our experiment, there is a random draw in each round to determine the size of the pot, which can be either large or small with equal probability. Crucially, only the first player knows the pot size. The first player can then give either a high or low share. A low share is considered to be an *equal* share if the pot is small or a *selfish* share if the pot is large. Because the second player does not observe the pot size, he cannot be certain whether a low share given by the first player is an equal or selfish share. We also measure the second player's reported level of anger with the first player.

Another aim of the experiment is to discriminate between two particular theories of intrinsic motivation. One is Levine's (1998) model of altruism and spitefulness, and the other is Akerlof's (2015) model of one's duty to comply with the norm. It would be difficult to test their specific predictions if there exist external incentives for punishment. For instance, in a typical multi-round game, a player might want to build a reputation of being a tough punisher in early rounds in order to influence the other player to give more in future rounds, and these future benefits become an external incentive. To isolate the intrinsic motivation for punishment, we do not inform the first player how much he is punished during the multi-round game, and the second player is also aware of this. This design resembles situations where people criticize or accuse others behind their back as a form of secret punishments.

We build a simple model to compare different theoretical predictions.

[1]The article by Trounstine (2014) gave many examples demonstrating how 'a true tragedy, driven by a media frenzy ... leads to bad public policy'. One such example is the 'Three Strikes and You're Out' law, a form of graduated sanctions. It was enacted by California voters in response to two horrific murders in the 1990s.

Because of the uncertainty about the pot size, the second player's belief about the first player's type might be updated by Bayes' rule based on the observed sharing decisions. The standard self-interest assumption predicts that no one will punish in any round because punishing is costly. Most theories of social preferences such as inequality aversion (e.g. Fehr & Schmidt, 1999) and reciprocity (e.g. Dufwenberg & Kirchsteiger, 2004) can predict positive punishment when there is a high likelihood of unequal outcomes or of unkind intention, but they cannot predict gradual increments of punishments. The reason is that these punishments are used to reduce inequality or to reciprocate unkind intention, rather than to proportionally mitigate the anger provoked.

Both Levine (1998) and Akerlof (2015) can explain why the severity of punishment might increase in an gradual manner. Levine proposes that people care about whether their partner is an altruistic or spiteful person. Hence, people can be kind to a partner who is believed to be altruistic; people can be unkind to one who is believed to be spiteful. Levine predicts that people will punish in proportion to the belief updated by Bayes' theorem about their partner's selfishness, regardless of whether the partner gives a high or low share in the current round. On the other hand, Akerlof assumes that some people have a norm, such as an equal sharing norm, and expect their partner to comply with the norm by giving an equal share. Anger is then provoked by the extent to which the partner fails to comply with the norm. Akerlof predicts that people will punish in proportion to the updated belief about their partner's selfishness, when their partner gives a low share in the current round; otherwise, they do not punish at all because giving a high share is in compliance with the norm.

Our results show that the first player's history of sharing decisions significantly affects the second player's current punishment level, even after controlling for the current sharing decision. More specifically, punishments escalate with the proportion of low shares the first player gives in previous rounds. This provides the first experimental evidence of escalating punishments.

Next, we test whether punishment depends solely on the belief updated by Bayes' rule about the first player's type (as in Levine) or punishment depends on both the updated belief and the current sharing decision (as in Akerlof). We find stronger support for Akerlof's model than for Levine's model. Therefore, anger and punishments escalate with the updated belief about the partner's selfishness only when the partner's current action fails to comply with

the sharing norm. When the partner's current action complies with the sharing norm, the anger accumulated in previous rounds is mitigated and triggers no harsher punishment.

Another contribution is to identify subjects who are likely to hold sharing norms, which influence both how they behave and how they expect others to behave. Importantly, not everyone holds the same expectation about how much others should give. Suppose a person expects his partner to give an equal share. When his partner's current action complies with the expected sharing norm, no punishment is deserved, even if he believes his partner is a selfish person due to past actions. Indeed, we find in our sample about one third of subjects behave in this way and punish only when there is non-compliance with the norm in the current action. Although about a quarter of subjects never punish, they still display escalating anger as if they expect others to comply with a sharing norm. Moreover, if people believe that a sharing norm exists, they should expect not only others but also themselves to give. Since all subjects play a game as the first player and another as the second player, we can and do find that subjects who tend to punish are more likely to give than those who never punish.

1.2 Literature Review

This study contributes to the experimental literature on testing recently developed theories. There is a large body of empirical evidence on the impact of fairness considerations on economic decisions, such as refusing to cut wages during a recession (e.g. Kahneman *et al.* , 1986; Bewley, 1999) and enforcing contracts and social norms (e.g. Fehr & Gächter, 2000b; Camerer & Thaler, 1995). This has inspired a number of new theories that incorporate nonstandard preferences, such as altruism (Levine, 1998), inequality aversion (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000), reciprocity (Rabin, 1993; Dufwenberg & Kirchsteiger, 2004), fairness intentions coupled with distributional equality (Falk & Fischbacher, 2006), and anger over non-compliance (Akerlof, 2015). These new theories not only help explain the empirical departures from the predictions of standard economic theory but also furnish many testable hypotheses.

This research builds on the work of designing new experiments to test the fine differences between competing hypotheses (Charness & Rabin, 2002;

Falk *et al.* , 2005, 2008). Our experiment is designed to investigate how concerns for fairness or anger over non-compliance accumulate over time. In keeping with the previous related literature, our goal is to provide concrete evidence to guide the effort of developing new economic theories.

Our experiment is closely related to Ostrom *et al.* (1992), who studied the impact of sanctioning opportunity in a multi-round common pool resource game. In their experiment, the subjects could only choose to impose a fixed fine on others or not, so it was not possible to impose different levels of punishments. There were also material incentives for punishing because punishments in early rounds could affect other subjects' decision to cooperate in future rounds. To rule out such material incentives in our experiment, we deliberately conceal from the first player the feedback about how much he was punished in previous rounds until the end of the experiment.

A related seminal study by Fehr & Gächter (2000a) used a different design to rule out future material incentives for punishment. Their multi-round public goods experiment changed the composition of group members from round to round so that players could not build individual reputations, but the results still indicated a widespread willingness to punish. Fehr & Gächter (2002) further showed the proximate cause for this kind of punishment is anger. Their design did not provide the possibility of identifying other members' history of contributions to the public goods. Our experiment in contrast does allow the second player to identify their partner's history of sharing decisions. The purpose is to examine how a history of sharing decisions can affect current anger and punishment.

Finally, our study contributes to the experimental literature that examines the individual heterogeneity in inflicting costly punishments. Anderson & Putterman (2006) found three types of punishers in a public goods game. One type virtually never punished, another type punished in response to the extent of free riding, and the third type was called indiscriminate punishers because they punished aggressively at all levels of free riding. Carpenter (2007) also found support for this classification[2] and suggested that some punishers follow cultural norms because they punish despite paying a high cost of inflicting punishment. However, there is no direct evidence on this normative story. Our data provide deeper analysis by examining the determinants of

[2]Carpenter (2007) further divided their subjects by whether they were contributors or free riders themselves.

anger level for different types of punishers. Specifically, subjects who never punish have a similar pattern of escalating anger as other subjects who punish primarily when their partner gives a low share in the current round. This finding suggests that these two groups of subjects are likely to hold a similar sharing norm that can trigger their anger but have different tolerance for the cost of punishing. In contrast, a third group of subjects who indiscriminately punish their partner, who may give a high or low share in the current round, display no clear pattern of anger with regard to their partner's sharing history.

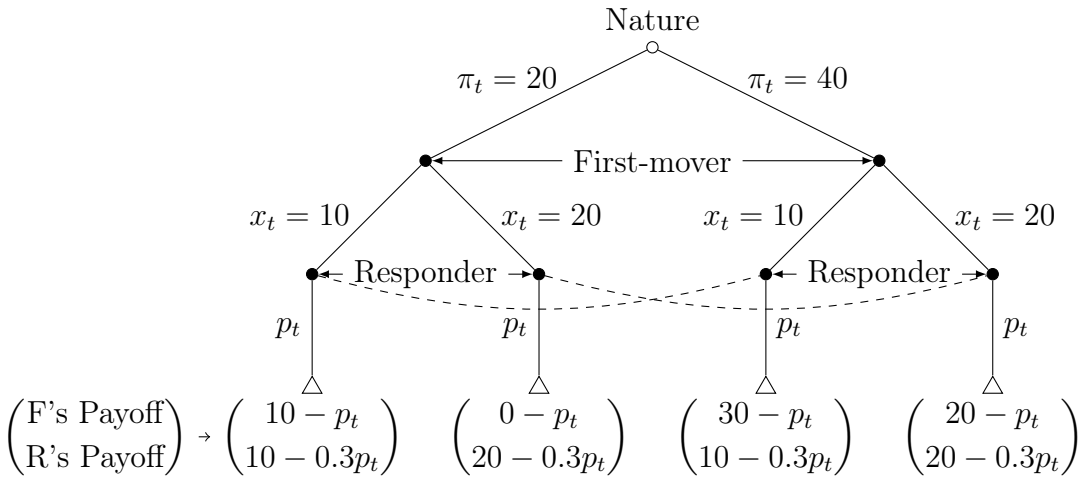
1.3 Experimental Design

The experimental design places subjects into groups of two, and each group plays a seven-round game. Within a group, one subject is assigned the role of the 'first-mover', and the other the role of the 'responder'. Each round $t \in \{1, 2, ..., 7\}$ consists of three stages. Let the pot size at round t be $\pi_t \in \{20, 40\}$. In the first stage, a pot of money is randomly drawn to be large (40) or small (20) with equal probability. In the second stage, the first-mover can choose to give a share of x_t to the responder and keep the rest of the pot for himself, $\pi_t - x_t$. The share x_t can be either high (20) or low (10). In the third stage, the responder observes how much the first-mover gives him but not the size of the pot.[3] The responder can then choose how much to punish the first-mover. Let the punishment unit be denoted by $p_t \in \{0, 1, 2, ..., 10\}$. Every unit of punishment reduces the first-mover's earnings by one unit and simultaneously costs the responder 0.3 units. The utility function at round t for the first-mover is $u_{F,t} = \pi_t - x_t - p_t$, and the utility function for the responder is $u_{R,t} = x_t - 0.3p_t$. Figure 1.1 presents the experimental procedure at round t. At the end of the seventh round the subjects switch roles and repeat the same game but are matched with a different partner.

The first-mover could not receive any feedback about how much he or she was punished by the responder in previous rounds, and this is common knowledge to both players. This design excludes *strategic* punishments, which are used to affect the first-mover's sharing behaviour in future rounds. Since there are no material benefits from using punishments in this game, the sub-

[3] Güth *et al.* (1996) uses a similar two-level ultimatum game with incomplete information and finds that better informed parties try to behave in a way that others cannot be sure that they are greedy. Thus, we can expect that the first-mover in our experimental design is likely to take advantage of the responder.

Figure 1.1: The Experimental Procedure at Round t.



jects' use of sanctions must suggest the presence of intrinsic motivation. This design feature is important for ensuring a clean test of Levine's notion of altruism and spitefulness and of Akerlof's concept of one's duty to comply with the norm.

To elicit punishment decisions, we apply the strategy method in our experiment, which requires the responder to provide a response for each feasible action of the first-mover before he is informed of the first-mover's actual choice. Specifically, we collect one punishment decision for the case when the first-mover gives a low share of 10 and another for the case when the first-mover gives a high share of 20. After making these two contingent punishment decisions, the responder is informed that the first-mover actually gives a high or low share. Consequently, only the punishment decision in that specific case affects payoffs. The entire procedure is carefully explained to the subjects in the instructions provided both on paper and on the computer screens, along with examples of screens which would be used in the experiment. Following that, we ask the subjects several questions and then show the answers to enhance their comprehension of the procedure.

In addition to examining the impact of the first-movers' sharing behaviour on the responders' punishment, we further investigate the responders' reported anger with the first-movers. If anger is the proximate cause of punishment, both punishment and anger should have similar patterns in responding to the sharing decisions. At the end of each round, we ask the responders to rate how angry they are with their partner on a seven-point scale ranging from

1 (*not at all*) to 7 (*very much*).

Since the responders cannot observe the pot size, we suspect that they would gradually form a belief about their partner's type through the observed sharing behaviour. To understand how this subjective belief might be formed, the responders are asked to estimate the likelihood of their partner giving a low share if the pot is large in the next round.[4] This variable is named the *likelihood of selfish share* and can take values from 0% to 100%.

All subjects in our experiment were students or staff from the University of Warwick, who were recruited online through the Warwick Research Participation System. The experiment was programmed in z-Tree (Fischbacher, 2007) and conducted in the computer laboratory at Warwick Business School in June and October 2014. A typical session lasted less than an hour, which included paying the subjects individually. Subjects received experimental tokens throughout the experiment, and each token was converted into £0.01. Thus, every decision counts, and the subjects each earned £3.85 on average.

1.4 Model and Predictions

We build a simple model to compare different theories of fairness and anger. Suppose there is a continuum of types distributed uniformly on $[0, 1]$ for the first-movers. A first-mover of type $q \in [0, 1]$ behaves *selfishly* with probability q and behaves *fairly* with probability $1 - q$. Behaving fairly is to give an equal share of the pot, whether large or small, and behaving selfishly is to give a low share regardless of the pot size. Hence, type q can be considered as a proxy of the degree of the first-mover's selfishness.

This distributional assumption on q reflects a likely scenario where even a fair-minded first-mover might be tempted to give selfishly occasionally, especially when, by chance, large pots are drawn more frequently than expected, because the responder cannot observe the actual pot size. Moreover, Güth *et al.* (1996) show that better informed people are inclined to "pretend fairness" when there is incomplete information. In our experiment, we also find that a substantial number of people behave fairly only some of the time. See Appendix for more detail.

[4] These elicited estimates were not financially incentivized. Gächter & Renner (2010) find that incentivizing beliefs can bias the levels of contribution in public goods experiments, although the stated beliefs are closer to the actual contributions of other group members.

The responder does not observe the first-mover's type. Given the assumption of uniform distribution, the prior expectation about the first-mover's type is $E(q) = 0.5$. The responder can update the posterior distribution of q at round t by Bayes' theorem according to the partner's sharing decisions up to and including round t.

The responder can observe in each round whether the first-mover gives a high or low share. However, since the pot size is unknown to the responder, he cannot be sure whether a low share ($x_t = 10$) is an equal share (if $\pi_t = 20$) or a selfish share (if $\pi_t = 40$). Let $B(t)$ be a Binomial random variable at round $t \in \{1, 2, ..., 7\}$ that represents the number of low shares the first-mover gives to the responder out of t rounds. The random variable $B(t)$ can take values from zero to t low shares. The probability of observing a low share from a first-mover of type q is $\Pr(x = 10|q) = \frac{1}{2} + \frac{q}{2}$, where the first term is the fifty percent chance of the pot being of a small size and the second term is the chance of the pot being of a large size and the first-mover behaving selfishly. After observing k number of low shares out of t rounds, the posterior probability density function of q at round t is $f(q|B(t) = k) = \frac{(1+q)^k(1-q)^{t-k}}{\int_0^1 (1+q)^k(1-q)^{t-k}dq}$. Let $\hat{q}(t, k) \equiv E(q|B(t) = k)$ be the expected value of the first-mover's selfishness after receiving k number of low shares out of t rounds.

The standard self-interest assumption predicts no one punishes because punishment incurs a cost and yields no material benefits to the punishers. Fehr & Schmidt's (1999) inequality aversion model and Dufwenberg & Kirchsteiger's (2004) reciprocity model can both predict the use of punishment as a means to reduce inequality or to reciprocate ill intention. However, they cannot account for the possibility of gradual increments of punishments. The reason is that the optimal level of punishment is not chosen in direct proportion to the belief updated by Bayes' rule about the partner's selfishness. Thus, most theories are incapable of predicting escalating punishments.

In contrast, Levine's (1998) model of altruism and spitefulness can provide an explanation for why the severity of punishment might increase in proportion to the belief updated by Bayes' rule about first-mover's selfishness. Levine (1998) proposes that people care whether their partner is an altruistic or spiteful person. Consequently, people can be kind to a partner who is believed to be altruistic; people can be unkind to one who is believed to be spiteful. Based on this insight, we assume that anger is a function of $\hat{q}(t, k)$, the belief about the first-mover's selfishness at round t. We can write the

responder's utility function at round t in the following functional form:

$$u_{R,t} = x_t - 0.3p_t - \frac{L\left(\hat{q}\left(t,k\right)\right)}{p_t}, \tag{1.1}$$

where $L\left(\hat{q}\left(t,k\right)\right)$ is Levine's function of anger, and $L\left(0\right) = 0$, $L\left(y\right) > 0$ and $L^{'}\left(y\right) > 0$ for $y \in (0,1]$. The last term of the utility function captures the motivation to punish in proportion to the anger provoked by the updated belief about first-mover's selfishness. The optimal punishment at round t in Levine's model is

$$p_t^L = \sqrt{\frac{L\left(\hat{q}\left(t,k\right)\right)}{0.3}}.$$

Levine's prediction: the level of punishment at round t increases in $\hat{q}\left(t,k\right)$, the updated belief about first-mover's selfishness after observing k low shares out of t rounds.

Another theory of anger is Akerlof's (2015) model of norm compliance. Akerlof assumes that subjects can hold a sharing norm and expect their partner to comply with the norm. Anger is then provoked by the extent to which the partner fails to comply with it. Suppose the responders hold an equal sharing norm that requires people to give $x_t = \pi_t/2$. Because π_t is unknown to the responder, he forms an expectation about the pot size, $E\left(\pi_t|x_t,\ \hat{q}\left(t,k\right)\right)$ based on the share at round t and the updated belief about first-mover's selfishness. The expected level of the first-mover's non-compliance with the equal sharing norm is $\left(\frac{E(\pi_t|x_t,\ \hat{q}(t,k))}{2} - x_t\right)^+$, where $\left(z\right)^+ = \max\left\{z,0\right\}$. The extent to which the first-mover fails to comply with the equal sharing norm in giving x_t determines the extent to which the responder is provoked to anger. Thus, we can write Akerlof's function of anger at round t as

$$A\left(\left(\frac{E\left(\pi_t|x_t,\ \hat{q}\left(t,k\right)\right)}{2} - x_t\right)^+\right), \tag{1.2}$$

where $A\left(0\right) = 0$, $A\left(y\right) > 0$ and $A^{'}\left(y\right) > 0$ for $y \in (0,5]$. If a first-mover gives a high share of $x_t = 20$ at round t, there will be no anger because a high share is either an equal or favourable share to the responder. If a first-mover gives a low share of $x_t = 10$ at round t, a responder's anger might be provoked and increase as the expected level of first-mover's non-compliance with the norm

rises. This expected non-compliance,

$$\frac{E\left(\pi_t | x_t = 10,\ \hat{q}\left(t,k\right)\right)}{2} - 10 = \frac{10\hat{q}\left(t,k\right)}{\left(1+\hat{q}\left(t,k\right)\right)},$$

is increasing in $\hat{q}\left(t,k\right)$, meaning that the more selfish a first-mover is believed to be, the more likely the first-mover would have given a low share out of a large pot. The responder's utility function is the same as equation (1.1) except that Levine's anger function is replaced by Akerlof's anger function in (1.2). Thus, the optimal punishment in Akerlof's model is

$$p_t^A = \begin{cases} \sqrt{\frac{A\left(\frac{10\hat{q}(t,k)}{(1+\hat{q}(t,k))}\right)}{0.3}}, & \text{if } x_t = 10 \\ 0, & \text{if } x_t = 20 \end{cases}.$$

Akerlof's prediction: the level of punishment at round t depends on both $x_t \in \{10, 20\}$, whether a first-mover gives a high or low share at round t, and $\hat{q}\left(t,k\right)$, the updated belief about the first-mover's selfishness after observing k low shares out of t rounds. When the share at round t is low, punishment increases in $\hat{q}\left(t,k\right)$. When the share at round t is high, no punishment is used.

Both Levine and Akerlof can also predict the behaviour of not punishing if $L(y)$ and $A(y)$ are assumed to be zero for all levels of anger. Otherwise, these models predict that punishment is a function of $\hat{q}\left(t,k\right)$, but they differ in the impact of $\hat{q}\left(t,k\right)$ on punishment when the first-mover gives a high share at round t. This key difference allows us to test econometrically which model can explain the punishment data better.

1.5 Empirical Strategy

We first investigate the impact of the first-movers' past and current sharing decisions on the responders' punishment decision at round t. We estimate the following panel data regression model:

$$p_{it} = \beta_0 + \beta_1 (History\ of\ low\ shares)_{it} + \beta_2 (Low\ share)_{it} + \alpha_i + \delta_t + \rho_i \delta_t + \epsilon_{it}, \tag{1.3}$$

where the dependent variable p_{it} is the punishment unit chosen by responder i at round t. The variable *History of low shares* at round t is the number of low shares i's first-mover gives in previous rounds divided by the number of

previous rounds $(t-1)$. Since the observations in the first two rounds have only one previous round or none, we include observations that have at least two rounds of history, that is, observations in the last five rounds. The variable *Low share* is a dummy indicating that a first-mover gives a low share at round t, instead of a high share. We include subjects fixed effects α_i as controls for individual-specific characteristics and round fixed effects δ_t as controls for each round of the experiment. To control for the effect of role reversal, we also include additional round fixed effects $\rho_i\delta_t$ for subjects who played the role of the responder after playing the role of the first-mover ($\rho_i = 1$).

The panel data fixed effects model captures within-subject variations, while removing any between-subject variation. This is an appropriate model for this experiment because our objective is to study how individuals respond to the history of interacting with their partner over time. A detailed discussion by Charness *et al.* (2012) on between-subject and within-subject designs also pointed out that although there are potential confounds in multi-round games, using fixed effects model can achieve consistency and is often able to test more complex hypothesis for which between-subject design is not feasible.

Second, we test two theoretical predictions by Levine (1998) and Akerlof (2015) regarding the role of the belief updated by Bayes' rule about first-mover's selfishness. Given the assumption of uniform distribution of first-mover's types, we can compute all possible beliefs of $\hat{q}\,(t,k)$, which is a function of k number of low shares and t rounds. To test the effect of this Bayesian belief on punishment, we estimate the following fixed effects panel data regression model:

$$\begin{aligned} p_{it} = \gamma_0 + \gamma_1 \hat{q}(t,k)_{it} \times (High\ share)_{it} + \gamma_2 \hat{q}(t,k)_{it} \times (Low\ share)_{it} \\ + \alpha_i + \delta_t + \rho_i\delta_t + \epsilon_{it}. \end{aligned} \tag{1.4}$$

Note that a separate variable for the sharing decision at round t is not needed in this equation because $\hat{q}\,(t,k)$ already takes it into account. The coefficient γ_1 is interpreted as the extent to which the Bayesian belief about the first-mover's selfishness affects the level of punishment when a first-mover gives a high share at round t. The coefficient γ_2 indicates the impact of the Bayesian belief about first-mover's selfishness on the level of punishment when a first-mover gives a low share at round t. According to Levine's prediction that people punish solely based on the updated belief about their partner's type,

Levine's hypotheses are:

$$\gamma_1 > 0 \text{ and } \gamma_2 > 0.$$

According to Akerlof's prediction that people punish according to the updated belief about their partner's selfishness only when their partner gives a low share at round t, Akerlof's hypotheses are:

$$\gamma_1 = 0 \text{ and } \gamma_2 > 0.$$

1.6 Results

Table 1.1 provides the summary of statistics for key variables by rounds. 122 subjects participate in the experiment. In each round, the pot size is drawn to be large about fifty percent of the time, but the percentage of first-movers giving a low share is at least 73%. This percentage difference implies that a substantial number of first-movers give a low share even when the pot is large. About 50% to 59% of responders choose to punish in a given round, and about 47% to 66% of responders rate that they are at least *a little* angry with their partner (Anger level ≥ 2). The responders on average estimate that about 60% to 70% of the time the first-mover would give a low share if the pot is large in the next round.

The results are divided into three subsections. The first subsection illustrates how the first-movers' past and current sharing decisions might influence the responders' punishment and then provides statistical analysis on the determinants of punishment. To better understand the motivations behind punishments, we also examine two self-report measures: the anger level with the partner and the estimated likelihood that the partner would give selfishly. Next, we test theoretical predictions on the effect of Bayesian updating about the first-mover's selfishness on punishment. Lastly, we show three identifiable patterns of punishment and anger in the sample.

1.6.1 The History of Sharing Behaviour

To study the effect of the history of sharing behaviour on punishment at round t, we create a variable called *history of low shares* to represent the past sharing decisions up to $t-1$ round, apart from the sharing decision at round t. A history of low shares at round t is defined as the number of low shares a

Table 1.1: Summary Statistics by Rounds

	Round t						
Variables	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**
Pot size is large	50.8%	51.6%	49.2%	57.4%	42.6%	57.4%	45.1%
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Giving a low share	77.0%	83.6%	77.0%	73.0%	82.0%	80.3%	91.2%
	(0.42)	(0.37)	(0.42)	(0.45)	(0.39)	(0.40)	(0.28)
Punishment unit	2.52	2.58	3.31	3.21	3.25	3.34	3.62
	(3.09)	(3.37)	(3.77)	(3.58)	(3.66)	(3.96)	(3.93)
Punishment frequency	54.1%	50.0%	57.4%	56.6%	59.0%	53.3%	57.4%
	(0.50)	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)	(0.50)
Anger level	1.86	2.14	2.37	2.53	2.70	2.82	3.10
	(1.19)	(1.39)	(1.57)	(1.79)	(1.81)	(1.87)	(2.06)
Anger frequency	46.8%	53.3%	57.4%	54.9%	63.1%	63.1%	66.4%
	(0.50)	(0.50)	(0.50)	(0.50)	(0.48)	(0.48)	(0.47)
Likelihood of selfish share	60.3%	62.2%	60.8%	64.9%	68.0%	68.0%	69.6%
	(0.25)	(0.26)	(0.28)	(0.26)	(0.25)	(0.29)	(0.28)
History of low shares		77.0%	80.3%	79.2%	77.7%	78.5%	78.8%
		(0.42)	(0.29)	(0.25)	(0.22)	(0.21)	(0.19)
Subjects	122						

Notes: In the first round, the pot size is large ($\pi = 40$) 50.8% of the time (as opposed to $\pi = 20$), and the first-movers give a low share ($x = 10$) 77.0% of the time (as opposed to $x = 20$). The responders choose on average 2.52 punishment units in the first round. Punishment frequency refers the percentage of subjects who choose a punishment that is at least one unit ($p \geq 1$). After observing the share a first-mover gives at round t, the responder indicates the level of anger on a seven-point scale ranging from 1 (*not at all*) to 7 (*very much*). Anger frequency refers the percentage of subjects who report at least *a little* angry with their partner (Anger level ≥ 2). The variable *likelihood of selfish share* refers to the responder's estimated likelihood that the partner would give a low share if the pot is large in the next round. The variable *history of low shares* at round t is defined as the number of low shares the responder received in previous rounds divided by the number of previous rounds $(t-1)$. Standard deviations are in parenthesis.

first-mover gives to a responder in previous rounds divided by the number of previous rounds $(t-1)$. Put simply, it is a fraction of low shares in previous rounds. This variable measures the degree of the lack of generosity the first-mover has shown to the responder in the past. Take the third round for example. Eighty subjects have a 100% history of low shares because there were two low shares in the first two rounds, thirty-six subjects who observed exactly one low share in the two previous rounds have a 50% history of low shares, and six subjects who observed no low shares have a 0% history of low

shares. Table 1.2 presents the distribution of the history of the low shares by rounds.

Table 1.2: Distribution of Responders Based on the History of Low Shares Received

History of low shares	Number of subjects					
	Round t					
	2nd	**3rd**	**4th**	**5th**	**6th**	**7th**
100%	94	80	62	44	39	35
83%						43
80%					53	
75%				57		
67%			47			28
60%					17	
50%		36		13		10
40%					8	
33%			10			4
25%				6		
20%					5	
17%						2
0%	28	6	3	2	0	0

Notes: A history of low shares at round t is defined as the number of low shares a responder receives from a first-mover in previous rounds divided by the number of previous rounds. In each round, the numbers in a column sum up to 122 subjects. Take the second round for example: 94 subjects have a 100% history of low shares because there was one low share in the first round, and 28 subjects have a 0% history of low shares because there was no low share. This variable represents the degree of the lack of generosity that a first-mover has shown to a responder in the past. The first round is omitted because it has no history.

Figure 1.2a illustrates the relationship between the first-movers' sharing decisions and the responders' punishment. Observations are divided into four categories based on the variable *History of low shares*, which is the number of low shares a responder received from a first-mover in previous rounds divided by the number of previous rounds. If a subject at round t received low shares less than 60% of the time in the past $t-1$ rounds, this observation falls into the first category with a label of '0%-'. If a subject at round t received low shares at least 60% and less than 80% of the time in the past $t-1$ rounds, this observation falls into the second category with a label of '60%-'. The left

section of the figure includes observations at round t in which the first-mover gives a high share. The right section is when the first-mover gives a low share at round t. Figure 1.2a suggests that the responders tend to inflict a harsher punishment when they receive a low share at round t than when they receive a high share. Moreover, a higher percentage of low shares in previous rounds seems to increase the severity of punishment, but formal statistical analysis is required to test this effect.

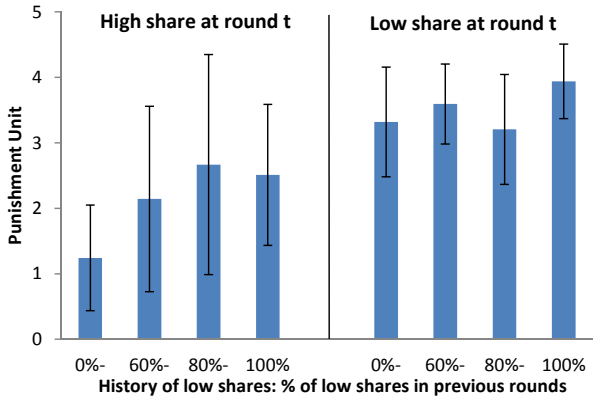
Some individuals may not be willing to pay the cost for inflicting punishment, though they still feel angry. To measure this feeling of anger, subjects are asked to rate how angry they are with their partner at the end of each round. We aim to check whether the history of low shares has the same impact on anger level as they would on punishment level. Figure 1.2b has the same categories as in Figure 1.2a except that the punishment variable is replaced by the anger variable. A comparison of the two sections of Figure 1.2b suggests that the responders tend to report a higher level of anger at round t when they receive a low share than when they receive a high share. Given that subjects receive a low share at round t as in the right section of Figure 1.2b, there seems to be a slight upward trend between the anger level and the percentage of low shares in previous rounds.

We are also interested in how the responders' estimated likelihood about their partner's future sharing behaviour is affected by their partner's past and current sharing decisions. Figure 1.2c uses the variable *Likelihood of selfish share*, which is the responder's estimated likelihood that the first-mover would give a low share if the pot is large in the next round. Figure 1.2c also suggests that a higher percentage of low shares in previous rounds seems to increase the responder's estimation about the likelihood of the first-mover sharing selfishly, especially when the first-mover gives a low share at round t.

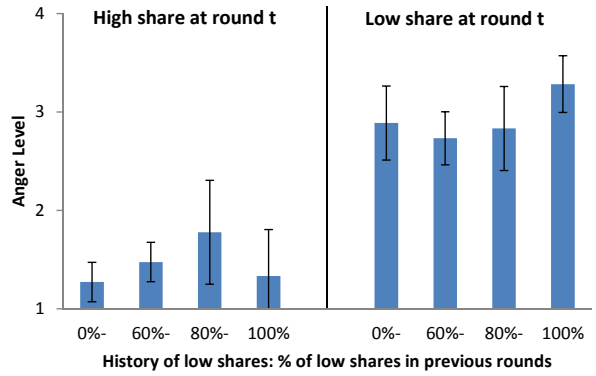
To provide statistical evidence that history matters for punishment, column 1 of Table 1.3 estimates the fixed effects panel data regression in equation (1.3). The point estimate on *Low share* in the second row means that the responders punish on average 1.8 units more when they receive a low share at round t than when they receive a high share. In the first row, the point estimate on *History of low shares* is positive and significant at 1% level after controlling for the effect of receiving a low share at round t, subject fixed effects, round fixed effects and the interaction terms between the dummy for role reversal and the round dummies. Thus, a higher proportion of low shares that

Figure 1.2: The Effects of the First-movers' Past and Current Sharing Decisions on the Responders' Punishment Decisions, Feelings of Anger, and Estimates of How Selfish Their Partner Might Be in the Future

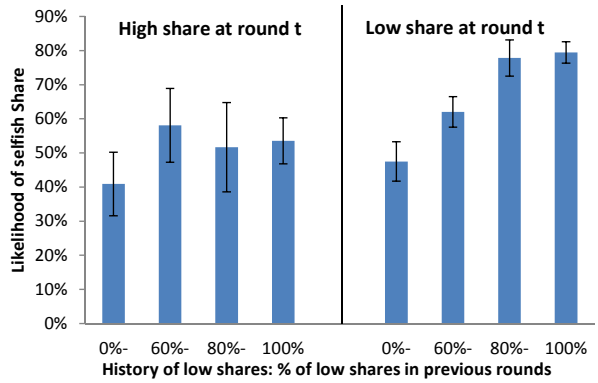
(a) Average Punishment



(b) Average Anger



(c) Average Estimate of How Selfish the First-mover Might Be in the Future



Notes: (a) The responder's punishment ranges from zero to ten units. A '0%-' history of low shares refers to observations in which the first-mover gives low shares less than 60% of the time in all previous rounds. A '60%-' history of low shares refers to at least 60% and less than 80% of the low shares in previous rounds. The left (right) section of the figure includes observations at round t in which the first-mover gives a high (low) share. When the first-mover gives a high share at round t, the number of observations for each of these four categories of the history of low shares from '0%-' to '100%': $N = 33, 21, 18, 45$. When the first-mover gives a low share at round t, the number of observations for each of these four categories: $N = 72, 128, 78, 215$. All 610 observations have at least two rounds of history of play for all subjects. Error bars are twice the standard error of the mean, approximating 95% confidence interval. (b) The responder's anger level ranges from 1 (*not at all*) to 7 (*very much*). (c) The responder's estimated likelihood that the first-mover would give a low share if the pot is large in the next round.

the first-mover gives in previous $t-1$ rounds causes the responder to increase the level of punishment at round t, holding constant whether the first-mover gives a high or low share at round t. This effect of history on punishment remains positive and significant at 1% level even if we include only observations that have at least three rounds of history of play for all subjects.

Column 2 of Table 1.3 uses the level of anger as the dependent variable in estimating equation (1.3). We find the same effect that the coefficient on *History of low shares* is significant and positive, suggesting that a higher

Table 1.3: The Impact of the First-movers' Sharing Behaviour on the Responders' Punishment Decisions, Feelings of Anger, and Estimates of How Selfish the Partner Might Be in the Future

Dependent Variable:	Punishment	Anger	Likelihood of selfish share
Independent Variables	(1)	(2)	(3)
History of low shares	2.870***	1.529***	0.354***
	(1.007)	(0.483)	(0.123)
Low share	1.816***	1.515***	0.0907**
	(0.466)	(0.183)	(0.0398)
Constant	-0.504	-0.883*	0.283**
	(0.990)	(0.503)	(0.127)
Subject FE	YES	YES	YES
Round FE	YES	YES	YES
Role Reversal × Round FE	YES	YES	YES
Observations	610	610	610
Subjects	122	122	122
R^2	0.073	0.271	0.082

Notes: Estimates from panel data fixed effects model. The dependent variable in column 1 is the level of punishment ranging from zero to ten units. The dependent variable in column 2 is the reported anger level with the partner, and the scale is adjusted to take values from zero (*not at all*) to six (*very much*). The dependent variable in column 3 is the responder's estimated likelihood that the partner would give a low share if the pot is large in the next round. The variable *Low share* is a dummy indicating that the first-mover gives a low share at round t. The variable *History of low shares* is defined as the number of low shares a first-mover gives in previous $t-1$ rounds divided by $t-1$. All observations have at least two rounds of history of play for all subjects. Robust standard errors in parentheses are clustered at the subject level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

proportion of low shares that a first-mover gives in previous rounds causes a responder to report a higher level of anger. In column 3, the dependent variable is the responders' estimated likelihood that the first-mover would give a low share when the pot is large in the next round. Again, we find that the history of low shares significantly increases the responders' estimated likelihood of how selfish their partner might be in the future.

1.6.2 A Test of Two Theories: Spitefulness and Norm Compliance

The evidence so far suggests that the responders base their punishment decisions not only on how much the first-mover gives at round t but also all sharing decisions in previous $t-1$ rounds. This is true also for how angry the responders report to be and for how much they estimate about their partner's future selfish behaviour. Now we further investigate whether these effects of past and current sharing decisions can be explained by the responders updating their belief about the first-mover's selfishness by Bayes' rule. The key variable is $\hat{q}\,(t,k)$, which is the expected belief about the partner's probability of behaving selfishly after observing k number of low shares out of t rounds. Table 1.4 tabulates all possible values of $\hat{q}\,(t,k)$ by t rounds and by k number of low shares.

Our aim is to econometrically test two theoretical predictions regarding the effect of $\hat{q}\,(t,k)$ on punishment. This effect is examined separately depending on whether the first-mover gives a high or low share at round t. When the first-mover gives a low share at round t, both Levine (1998) and Akerlof (2015) predict that $\hat{q}\,(t,k)$ will have a positive effect on punishment, that is, $\gamma_2 > 0$ in equation (1.4); this common hypothesis states that the responders choose a harsher punishment if they have a stronger belief about the partner's selfishness based on Bayes' rule. When the first-mover gives a high share at round t, Levine predicts $\gamma_1 > 0$ while Akerlof predicts that $\gamma_1 = 0$. Thus, the sign and significance of γ_1 allow us to discriminate between the two theories.

Column 1 of Table 1.5 tests these predictions by estimating the fixed effects panel data regression in equation (1.4). The estimate on γ_2 is positive and significant ($p = 0.040$), suggesting that when the first-mover gives a low share at round t, the responders tend to increase the level of punishment in proportion to the extent of the Bayesian belief about their partner's selfishness.

Table 1.4: The Responder's Belief Updated by Bayes' Rule About the First-mover's Selfishness

# of low shares out of t rounds: $B(t) = k$	The Bayesian belief about first-mover's selfishness at round t: $\hat{q}(t,k) = E(q\|B(t) = k)$						
	1st	2nd	3rd	4th	5th	6th	7th
0	0.333	0.250	0.200	0.167	0.143	0.125	0.111
1	0.556	0.375	0.280	0.222	0.184	0.156	0.136
2		0.607	0.418	0.313	0.247	0.203	0.171
3			0.653	0.462	0.347	0.273	0.223
4				0.694	0.504	0.383	0.302
5					0.728	0.544	0.419
6						0.757	0.581
7							0.781

Notes: The first-movers are assumed to have a continuum of types distributed uniformly on $[0, 1]$. A first-mover of type $q \in [0, 1]$ behaves *selfishly* with probability q (giving a low share regardless of the pot size) and behaves *fairly* with probability $1 - q$ (giving an equal share of the pot, whether large or small). The responder's prior expectation about the first-mover's selfishness is $E(q) = 0.5$. After observing k number of low shares out of t rounds, the responder's posterior expectation about the first-mover's selfishness is $\hat{q}(t, k)$.

In contrast, the estimate on γ_1 is not significantly different from zero ($p = 0.571$), implying that when the first-mover gives a high share at round t, a stronger Bayesian belief about the first-mover's selfishness does not lead to a harsher punishment. These results confirm that Akerlof's model of norm compliance fits the data better than Levine's model of spitefulness. Therefore, people are likely to evaluate their partner's current action against a certain sharing norm, in addition to evaluating their partner's selfishness based on past and current sharing decisions.

We also test these theoretical predictions with regard to the responder's level of anger. Column 2 of Table 1.5 uses the anger level as the dependent variable to estimate equation (1.4) . The results are similar to those in column 1 of Table 1.5. The estimate on $\hat{q}(t, k) \times High\ share$ is not significant ($p = 0.461$), while the estimate on $\hat{q}(t, k) \times Low\ share$ is positive and significant ($p < 0.001$). This further confirms Akerlof's model that anger is potentially triggered by the extent to which a current action fails to comply with the equal sharing norm. A stronger Bayesian belief about the first-mover's selfishness leads the responder to believe that a low share given by the first-mover is more likely to be a selfish share than an equal share.

Table 1.5: The Role of the Belief Updated by Bayes' Rule on the Responders' Punishment Decisions, Feelings of Anger, and Esimates of How Selfish the Partner Might Be in the Future

Dependent Variable:	Punishment	Anger	Likelihood of selfish share
Independent Variables	(1)	(2)	(3)
$\hat{q}(t,k) \times High\ share$	1.426	0.807	0.687***
	(2.510)	(1.091)	(0.239)
$\hat{q}(t,k) \times Low\ share$	4.093**	3.271***	0.700***
	(1.976)	(0.907)	(0.216)
Constant	1.433	0.356	0.311**
	(1.126)	(0.502)	(0.122)
Subject FE	YES	YES	YES
Round FE	YES	YES	YES
Role Reversal × Round FE	YES	YES	YES
Observations	610	610	610
Subjects	122	122	122
R^2	0.062	0.249	0.095

Notes: Estimates from panel data fixed effects model. The dependent variable in column 1 is the level of punishment ranging from zero to ten units. The dependent variable in column 2 is the reported anger level with the partner, and the scale is adjusted to take values from zero (*not at all*) to six (*very much*). The dependent variable in column 3 is the responder's estimated likelihood that the partner would give a low share if the pot is large in the next round. The variable $\hat{q}(t,k)$ is the responder's expected probability of the first-mover behaving selfishly after receiving k low offers out of t rounds. The variable $High\ share$ ($Low\ share$) is a dummy indicating that a first-mover gives a high (low) share at round t. All observations have at least two rounds of history of play for all subjects. Robust standard errors in parentheses are clustered at the subject level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Column 3 of Table 1.5 checks whether the belief updated by Bayes' rule can predict the responder's estimated likelihood that the first-mover would give a low share if the pot is large in the next round. Notably, the coefficient on $\hat{q}(t,k) \times High\ share$ is positive and significant at 1% level, suggesting that when the first-mover gives a high share at round t, the responder's estimated likelihood about the first-mover future selfish behaviour is still in direct proportion to $\hat{q}(t,k)$. Moreover, this coefficient is not significantly different from the positive coefficient on $\hat{q}(t,k) \times Low\ share$ ($p = 0.836$); thus, the respon-

ders estimate the likelihood of their partner's future selfish behaviour in a way that is similar to Bayesian updating, regardless of whether their partner gives a high or low share at round t.

A crucial point is that whether the current sharing decision has a mediating effect on the impact of $\hat{q}(t,k)$ on the dependent variable in Table 1.5. In column 3 of Table 1.5, there is no mediating effect because the estimated likelihood of the first-mover's future selfish behaviour is similar to the updated belief by Bayes' rule, which has already taken into account both past and current sharing decisions. In contrast, column 1 of Table 1.5 shows that the current sharing decision does have a significant mediating effect on the impact of $\hat{q}(t,k)$ on punishment decisions ($H_0 : \gamma_1 = \gamma_2$; $p = 0.027$). This finding is predicted by Akerlof's model of norm compliance, which pays more attention to the partner's current action than the belief about the partner's type.

1.6.3 Individual Heterogeneity in Punishment Decisions

Recent experiments found that the tendency to punish differs considerably across individuals and found support for identifying different types of punishers (Carpenter, 2007; Anderson & Putterman, 2006). In our experiment, we identify three basic groups of players based on two dimensions of contingent decisions. In one dimension which concerns the case when the first-mover gives a low share, players are sorted by the number of times they choose to punish their partner over the seven rounds in a game. On one end of this dimension, those who never punish in a game are defined as *non-punishers*.[5] Those who punish at least once for receiving a low share over the second rounds are called punishers. These punishers are then sorted in the second dimension, which concerns the case when the first-mover gives a high share, by the number of times they choose to punish their partner over the seven rounds in a game (see Appendix B for more detail). We identify those who punish more than once for receiving a high share as *excessive punishers*. The rest are *moderate punishers* because they choose zero punishment for receiving a high share at least six times out of seven rounds.

The distribution of these three groups of players is as follows: 26.2% are non-punishers, 32.0% are moderate punishers, and 41.8% are excessive punishers. Note that this classification is based on the frequency of punishing,

[5] In the case when their partner gives a high share, all non-punishers (except one) never punish their partner. This exception has only one positive punishment out of seven rounds.

not on the level of chosen punishment. Thus, excessive punishers do not necessarily inflict higher levels of punishment; they only punish more frequently than moderate punishers when they are given a high share. To understand their punishment behaviour, subjects were asked to explain why they made the decisions they made in a questionnaire at the end of the experiment. A few subjects appeared to be confused with the experiment, while others might want to revenge because they felt that they were taken advantage in previous rounds.

Table 1.6: Punishers Are More Likely to Give than Non-punishers

Dependent Variable:	Give high share $\in \{0, 1\}$	
	Logit	OLS
Independent Variables	(1)	(2)
Moderate punishers	0.097**	0.113***
	(0.038)	(0.041)
Excessive punishers	0.096***	0.111***
	(0.032)	(0.036)
Large pot size	0.178***	0.187***
	(0.031)	(0.031)
Constant		0.070
		(0.060)
Round FE	YES	YES
Role Reversal × Round FE	YES	YES
Observations	854	854
Subjects	122	122
R-squared		0.105

Notes: Column 1 reports the marginal effects for logit model, and column 2 reports OLS estimates. The dependent variable is a first-mover's decision to give a high share or not at round t. These first-movers are classified into three groups based on their punishment decisions when they play the role of the responder. The default group is *non-punishers*. The binary variable *Moderate punishers* indicates whether a first-mover is a moderate punisher. The binary variable *Excessive punishers* indicates whether a first-mover is an excessive punisher. The dummy variable *Large pot size* indicates whether the pot size at round t is large. Standard errors in parentheses are clustered at the subject level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Since every subject plays a game as the responder and another as the first-mover, we can assess whether these three identified groups behave differ-

ently when they are put in the role to give. We suspect that punishers, who might hold a sharing norm, are more likely to give a high share to others than non-punishers. Indeed, we find that 46.9% of non-punishers give a high share at least once out seven rounds, while this percentage is 82.1% for moderate punishers and 78.4% for excessive punishers. Both types are punishers are significantly more likely to give high shares than non-punishers (Fisher's exact test, $p = 0.001$).

Regression analysis also suggests that the motivation to punish is positively correlated with the motivation to give. Table 1.6 regresses a first-mover's decision to give a high share at round t on three binary variables, round fixed effects and interaction terms between a role reversal dummy and round dummies. The first variable *Moderate punishers* indicates whether a first-mover is identified as a moderate punisher when playing the role of the responder. The second variable *Excessive punishers* indicates whether a first-mover is an excessive punisher. The default group is thus non-punishers. The third variable *Large pot size* indicates whether the pot size at round t is large. Column 1 of Table 1.6 reports the marginal effects on giving a high share in logit model. The coefficients on *Moderate punishers* and *Excessive punishers* are both positive and significant, suggesting that both moderate and excessive punishers are about 9.7 percentage points more likely to give a high share than a non-punisher. The positive and significant coefficient on *Large pot size* means that the first-movers are 17.8 percentage points more likely to give a high share if the pot at round t is large, rather than small.

By definition, excessive punishers are more prone than moderate punishers to punish a partner who gives a high share. Moderate punishers behave in a way that is similar to Akerlof's prediction that people would refrain from inflicting punishment whenever their partner gives a high share in the current round. On the other hand, excessive punishers might punish a partner who gives a high share because of Levine's reasoning, that is, the partner is believed to be a selfish person. Table 1.7 tests these conjectures by estimating equation (1.4) separately for moderate punishers in column 1 and for excessive punishers in column 2. As expected, column 1 shows an insignificant coefficient of γ_1 and a significantly positive coefficient of γ_2 ($p = 0.037$), suggesting that moderate punishers follow Akerlof's model of norm compliance. Surprisingly, column 2 of Table 1.7 shows that both coefficients of γ_1 and γ_2 are insignificantly different from zero, suggesting that excessive punishers do not follow

Levine's model of spitefulness. On the contrary, the constant term ($p = 0.040$) in column 2 indicates that excessive punishers choose an average punishment as high as 4.3 units, even if their updated belief by Bayes' rule indicates that the first-mover is not a selfish person at all ($\hat{q}(t,k) = 0$). Despite high level and frequency of punishment, excessive punishers do not punish a partner more because the partner is more likely to be a selfish person.

Table 1.7: Heterogeneous Effects of Bayesian Updating on Punishment and Anger

Dependent Variable:	Punishment		Anger		
	Moderate Punishers	Excessive Punishers	Non-Punishers	Moderate Punishers	Excessive Punishers
Independent Variables	(1)	(2)	(3)	(4)	(5)
$\hat{q}(t,k) \times High\ share$	1.407	2.410	3.960*	2.252	-2.758*
	(4.916)	(4.540)	(2.003)	(1.612)	(1.389)
$\hat{q}(t,k) \times Low\ share$	8.547**	3.537	5.969***	5.243***	-0.157
	(3.963)	(3.605)	(1.762)	(1.342)	(1.033)
Constant	-1.193	4.314**	-2.264**	-0.087	2.388***
	(2.135)	(2.048)	(1.039)	(0.701)	(0.546)
Subject FE	YES	YES	YES	YES	YES
Round FE	YES	YES	YES	YES	YES
Role Rev. × Round FE	YES	YES	YES	YES	YES
Observations	195	255	160	195	255
Subjects	39	51	32	39	51
R^2	0.217	0.046	0.355	0.451	0.163

Notes: Estimates from panel data fixed effects model. For columns 1 and 2, the dependent variable is the level of punishment ranging from zero to ten units. For columns 3 to 5, the dependent variable is the responder's anger level, and the scale is adjusted to take values from zero (*not at all*) to six (*very much*). The variable $\hat{q}(t,k)$ is the responder's expected probability of the first-mover behaving selfishly after receiving k low offers out of t rounds. The variable $High\ share$ ($Low\ share$) is a dummy indicating that the first-mover gives a high (low) share at round t. All observations have at least two rounds of history of play for all subjects. Robust standard errors in parentheses are clustered at the subject level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Columns 3 to 5 of Table 1.7 estimate equation (1.4) while using the responder's anger level as the dependent variable for all three groups of subjects. We can estimate on the subsample of non-punishers because they could potentially report that they are angry with their partner even though they

Table 1.8: Heterogeneous Effects of Bayesian Updating on the Responders' Estimates of How Selfish the Partner Might Be in the Future

Dependent Variable:	Likelihood of selfish share		
	Non-Punishers	Moderate Punishers	Excessive Punishers
Independent Variables	(1)	(2)	(3)
$\hat{q}(t,k) \times High\ share$	0.358	1.534***	0.246
	(0.504)	(0.325)	(0.371)
$\hat{q}(t,k) \times Low\ share$	0.352	1.457***	0.330
	(0.517)	(0.283)	(0.301)
Constant	0.542*	-0.123	0.508***
	(0.307)	(0.156)	(0.167)
Subject FE	YES	YES	YES
Round FE	YES	YES	YES
Role Reversal × Round FE	YES	YES	YES
Observations	160	195	255
Subjects	32	39	51
R^2	0.069	0.254	0.091

Notes: Estimates from panel data fixed effects model. The dependent variable is the responder's estimated likelihood that the first-mover would give a low share if the pot is large in the next round. The variable $\hat{q}(t,k)$ is the responder's expected probability of the first-mover behaving selfishly after receiving k low offers out of t rounds. The variable $High\ share$ ($Low\ share$) is a dummy indicating that the first-mover gives a high (low) share at round t. All observations have at least two rounds of history of play for all subjects. Robust standard errors in parentheses are clustered at the subject level.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

never punish. One possible reason for not punishing is that non-punishers do not think others have a duty to give a high share, so no anger is triggered if receiving a low share. However, in column 3 of Table 1.7 the positive and significant coefficient on $\hat{q}(t,k) \times Low\ share$ ($p = 0.002$) suggests that non-punishers still report a higher level of anger in response to a stronger Bayesian belief about their partner's selfishness, when their partner gives a low share at round t. Notably, the coefficient on $\hat{q}(t,k) \times High\ share$ is positive and marginally significant ($p = 0.057$); thus, receiving a high share from the partner at round t does not necessarily stop non-punishers from reporting higher

levels of anger in response to a stronger Bayesian belief about the partners' selfishness. In contrast, this effect of $\hat{q}(t,k) \times High\ share$ on anger is not present with either moderate or excessive punishers in column 4 or 5 of Table 1.7. The reason for the difference between non-punishers and punishers might be that the use of punishment in a given round is able to mitigate the anger provoked in that round. Hence, if anger was provoked in previous rounds for non-punishers, this anger could be built up and remain high even when the partner gives a high share in the current round.

Another interesting difference between groups is found in the responder's estimated likelihood that their partner would give a low share if the pot is large in the next round. Table 1.8 uses this estimated likelihood as the dependent variable to separately estimate equation (1.4) for each of the three groups. Remarkably, only moderate punishers in column 2 tend to give a higher estimate that their partner might behave selfishly in response to a stronger Bayesian belief about the partner's selfishness. In contrast, both non-punishers in column 1 and excessive punishers in column 3 appear to be quite unresponsive in their estimated likelihood to the strength of Bayesian belief. This again confirms that excessive punishers do not think or behave in a way described in Levine's model of spitefulness.

In summary, we discover in our sample three distinct patterns of punishment and anger in response to the belief updated by Bayes' rule about the partner's selfishness. First, non-punishers tend to escalate their anger in spite of never choosing to punish. Second, moderate punishers tend to escalate anger and punishments only when their partner gives a low share. Third, the punishments and anger of excessive punishers display no escalating pattern, despite the high level and frequency of their punishment.

1.7 Conclusion

This chapter presents an experiment that explores the factors can explain how much people punish potential transgressors in a multi-round game. The design eliminates any material incentives for inflicting punishment, such as the benefits gained by building a reputation as a tough punisher. Hence, the motivation to punish must be something intrinsic. Our results provide the first experimental evidence that people tend to choose a harsher punishment in response to a higher proportion of low shares their partner gives in previous

rounds. Analysis of the responder's level of anger reveals a similar escalating pattern as in the responder's punishment decisions. This finding suggests that the human emotion of anger can have a cumulative effect on how people punish a repeat transgressor. Also, the escalating anger can potentially be one of the driving forces behind the widespread legislations of graduated sanctions.

Our data provide further evidence concerning the role of the responder's belief updated by Bayes' theorem about the first-mover's selfishness in punishment decisions. We find more support for the theory of norm compliance (Akerlof, 2015) that predicts people punish according to the extent to which the partner's action fails to comply with a sharing norm than for the theory of spitefulness (Levine, 1998) that predicts people punish a spiteful person regardless of the partner's current action. This evidence suggests that cumulative anger can be explained by the updated belief about how much an action fails to comply with the norm.

We also find that individuals differ in the extent of internalizing the sharing norm and identify three patterns of behaviour. First, a third of subjects punish an action that potentially fails to comply with the norm. Second, a quarter of subjects never punish yet may still get angry. Third, the rest punish indiscriminately a selfish or non-selfish partner. Moreover, internalizing a sharing norm affects both how one behaves and how one expects others to behave. Punishers are found to be more likely to give in compliance with the sharing norm than non-punishers. Therefore, future research, whether theoretical or empirical, should pay more attention to the extent of norm internalization across individuals.

This experiment not only demonstrates that emotions can influence economic decisions but also discriminates between different economic models of anger. More experimental works can be done to examine how anger evolves through time in response to receiving new information and the economic consequences of accumulated or mitigated anger.

1.8 Appendix A: Experimental Instructions

Welcome to this study of decision-making. The experiment will take about 60 minutes. The instructions are simple, and if you follow them carefully, you can earn up to £6.[6] You will be earning 'experimental currency units' (ECUs), which will be converted to pound sterling at the rate of 1 ECU = £0.01. All the money you earn is yours to keep, and will be paid to you, in cash, in private, after the experiment ends. Your confidentiality is assured.

You will be randomly matched with another participant. Your earnings will depend on your decisions and his/her decisions. You will not know their identity in any given decision period or after the experiment is over, and they will not know yours.

You will be assigned one of two roles: the first-mover or the responder. You will interact with the same participant for 7 periods. Each period has two stages.

In the first stage, the first-mover will be asked to split a pot of money between the two. The pot size will be either 20 or 40 ECUs. Each with 50% chance. The responder will not observe the exact pot size, but the first-mover will. The first-mover can only choose to share either 10 or 20 with the responder, while keeping the rest of the pot for himself. For example, if he shares 10 with the responder, he keeps 10 when the pot is 20 or keeps 30 when the pot is 40.

If the responder receives:	10	20
When the pot is 20, the first-mover keeps:	10	0
When the pot is 40, the first-mover keeps:	30	20

In the second stage, the responder will have the option to reduce the first-mover's earnings. The amount of reduction can be between 0 and 10 ECUs. Every 1 ECU reduction will cost the responder 0.3 ECUs. For example, if the responder reduces first-mover's earnings by 9 ECUs, it costs the responder (0.3)(9) = 2.7 ECUs. The first-mover will not observe how much his earnings have been reduced until the end of the experiment.

If first-mover gets a reduction of:	0	1	2	3	4	5	6	7	8	9	10
then it will cost the responder:	0	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3

[6] In three out of five sessions, the instructions say that 'you can earn up to £8'. This is due to miscalculation. In our analysis we still pool the five sessions together because the Kolmogorov-Smirnov test suggests that we cannot reject equality of the two distributions.

Before seeing the share in each period, the responder needs to indicate how much he would like to reduce the first-mover's earnings both for the case when the share is 10 and when the share is 20. Of course, only one of the two cases would have been selected by the first-mover, so only in this specific case the choice of reduction will affect earnings.

You will answer a few comprehension questions. These are purely for your understanding of the procedures, so your answers will not affect your payment.

Please raise your hand if you have any questions. Once the experiment begins, there will be no further discussion, and no communication of any kind among the participants is permitted.

1.9 Appendix B: Additional Analysis

In section 1.4, the assumption that the first-movers' types are uniformly distributed between zero and one implies that a large portion of people behave fairly some of the time and behave selfishly some of the time. To see if this is true in our experiment, we can look at the number of times a large pot is randomly drawn over seven rounds and the number of times a first-mover gives a low share conditional on the pot being large. Table 1.9 shows the distribution of first-movers based on these two variables. Only 6.6% of subjects behave fairly by giving a high share whenever the pot is large, corresponding to type $q = 0$. About 40% behave selfishly by giving a low share whenever the pot is large, corresponding to type $q = 1$. Over 50% of subjects fall into the category of $0 < q < 1$, rendering support for the assumption that many people are partially fair and partially selfish.

Table 1.9: Distribution of First-movers by the Number of Large Pots and the Number of Low Shares Given Large Pots (%)

	Number of Low Shares Given Large Pots							
Number of Large Pots	0	1	2	3	4	5	6	Total
1	0.8	3.3						4.1
2	1.7	6.6	7.4					15.7
3	0.8	5.0	14.0	9.9				29.8
4	2.5	0.8	7.4	5.0	9.1			24.8
5	0.8	0.8	1.7	4.1	4.1	9.1		20.7
6	0.0	0.0	0.8	1.7	0.0	0.8	1.7	5.0
Total	6.6	16.5	31.4	20.7	13.2	9.9	1.7	100.0

Notes: The row variable of the contingency table is the number of large pots that are randomly drawn over seven rounds. The column variable is the number of low shares chosen by the first-mover conditional on the pot being large. For example, in the case where three large pots were drawn out of seven rounds, 0.8% never gave a low share, 5.0% gave one low share, 14.0% gave two low shares, and 9.9% gave three low shares. One subject happened to observe zero large pots over seven rounds, so it was not possible to analyse how fair his decisions were, and thus he was excluded.

In section 1.6, we use panel data fixed effects model to estimate the determinants of subject's reported anger. The advantage is that it controls for all the unobserved individual heterogeneity, but it does not account for the ordinal nature of the dependent variable. To address this issue, we use random effects ordered logit estimation, and Table 1.10 reports these results.

Table 1.10: Random Effects Ordered Logit Model on Reported Anger

Dependent Variable: Anger	All	All	Non-Punishers	Moderate Punishers	Excessive Punishers
Independent Variables	(1)	(2)	(3)	(4)	(5)
History of low shares	2.648***				
	(0.956)				
Low share	4.253***				
	(0.541)				
$\hat{q}(t,k) \times High\ share$		-3.063	-2.722	-7.232	-4.931*
		(2.432)	(10.25)	(7.400)	(2.683)
$\hat{q}(t,k) \times Low\ share$		5.430***	12.96	9.546***	0.918
		(1.741)	(7.930)	(3.021)	(1.983)
Round FE	YES	YES	YES	YES	YES
Role Rev. × Round FE	YES	YES	YES	YES	YES
Observations	610	610	160	195	255
Subjects	122	122	32	39	51

Notes: Estimates are from random effects ordered logit model. The dependent variable is the reported anger level with the partner on a scale from one (*not at all*) to seven (*very much*). The variable $\hat{q}(t,k)$ is the expected probability of the first-mover behaving selfishly after receiving k low offers out of t rounds. The variable *High share* is a dummy indicating that the first-mover gives a high share at round t. The variable *Low share* is one minus *High share*. All observations have at least two rounds of history of play. Robust standard errors in parentheses are clustered at the subject level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

In section 1.6.3, we identify three types of punishers based on subjects' contingent punishment decisions. Table 1.11 shows a detailed distribution of responders by the number of times they choose to punish conditional on the two cases of giving low and high shares. According to the previous definition, the first row represents the non-punishers because they never punished when receiving a low share. Among those who punished at least once when receiving a low share, namely the second row onwards, the first two columns represent the moderate punishers, and the rest of the columns represent the excessive punishers.

Next, I present evidence that the order of playing a given role does not have a significant effect on the subjects' decisions. Table 1.12 first shows that the frequency of giving a low share is not significantly different between subjects who play the first-mover as the first role or the second role in all rounds except the second and the sixth rounds at 10% level. The second half of Table 1.12 shows that the punishment distribution in each round for subjects

Table 1.11: Distribution of Responders by the Number of Decisions to Punish Conditional on Low Share and on High Share (%)

# Punish Conditional on Low Share	Number of Decisions to Punish Conditional on High Share								
	0	1	2	3	4	5	6	7	Total
0	25.4	0.8							26.2
1	4.1	0.8							4.9
2	0.8	1.6	0.8						3.3
3	2.5	0.8	0.8	0.0					4.1
4	2.5	0.0	0.0	0.0	0.8				3.3
5	1.6	4.1	0.0	1.6	0.0	0.8			8.2
6	3.3	1.6	0.0	0.0	0.0	1.6	0.8	1.6	9.0
7	4.9	3.3	3.3	2.5	0.8	0.8	3.3	22.1	41.0
Total	45.1	13.1	4.9	4.1	1.6	3.3	4.1	23.8	100.0

Notes: The row variable of the contingency table is the number of decisions to punish over the seven rounds conditional on receiving a low share. The column variable is the number of decisions to punish over the same seven rounds conditional on receiving a high share. Subjects in the first row are non-punishers because they never punished when receiving a low share. Among those who punished at least once when receiving a low share, the first two columns constitute moderate punishers and the rest of the columns constitute excessive punishers.

whose first role is the responder is essentially the same as the distribution for subjects whose second role is the responder.

Table 1.12: No Evidence of Order Effects on Both Giving and Punishment Behaviour

	Give Low Share		p-value
	First Role is	Second role is	Chi-square
Round	First-mover	First-mover	Test
1st	73.8%	80.3%	0.39
	(0.44)	(0.40)	
2nd	77.0%	90.2%	0.05
	(0.42)	(0.30)	
3rd	80.3%	73.8%	0.39
	(0.40)	(0.44)	
4th	77.0%	68.9%	0.31
	(0.42)	(0.47)	
5th	80.3%	83.6%	0.64
	(0.40)	(0.37)	
6th	88.5%	72.1%	0.02
	(0.32)	(0.45)	
7th	95.1%	88.5%	0.19
	(0.22)	(0.32)	

	Average Punishment		p-value
	First Role is	Second role is	Mann-
Round	Responder	Responder	Whitney Test
1st	2.48	2.56	0.62
	(0.41)	(0.38)	
2nd	2.95	2.21	0.31
	(0.47)	(0.39)	
3rd	3.69	2.93	0.37
	(0.51)	(0.45)	
4th	3.21	3.21	0.87
	(0.44)	(0.48)	
5th	3.54	2.97	0.37
	(0.48)	(0.46)	
6th	3.31	3.38	0.78
	(0.51)	(0.51)	
7th	4.15	3.10	0.13
	(0.52)	(0.47)	

Notes: Standard errors are in parenthesis. The null hypothesis of a chi-square test is that playing the first-mover as the first role or the second role has no effect on subjects' decisions to give a low share in a given round. The Mann-Whitney test hypothesizes that the punishment distribution is identical for the two groups of responders.

1.10 Appendix C: Screenshots

Figure 1.3: This screenshot displays what a first-mover might see in a given round. The pot is randomly drawn to large (40 ECUs) or small (20 ECUs), and the first-mover decides to give a low share (10 ECUs) or high share (20 ECUs) to the responder.

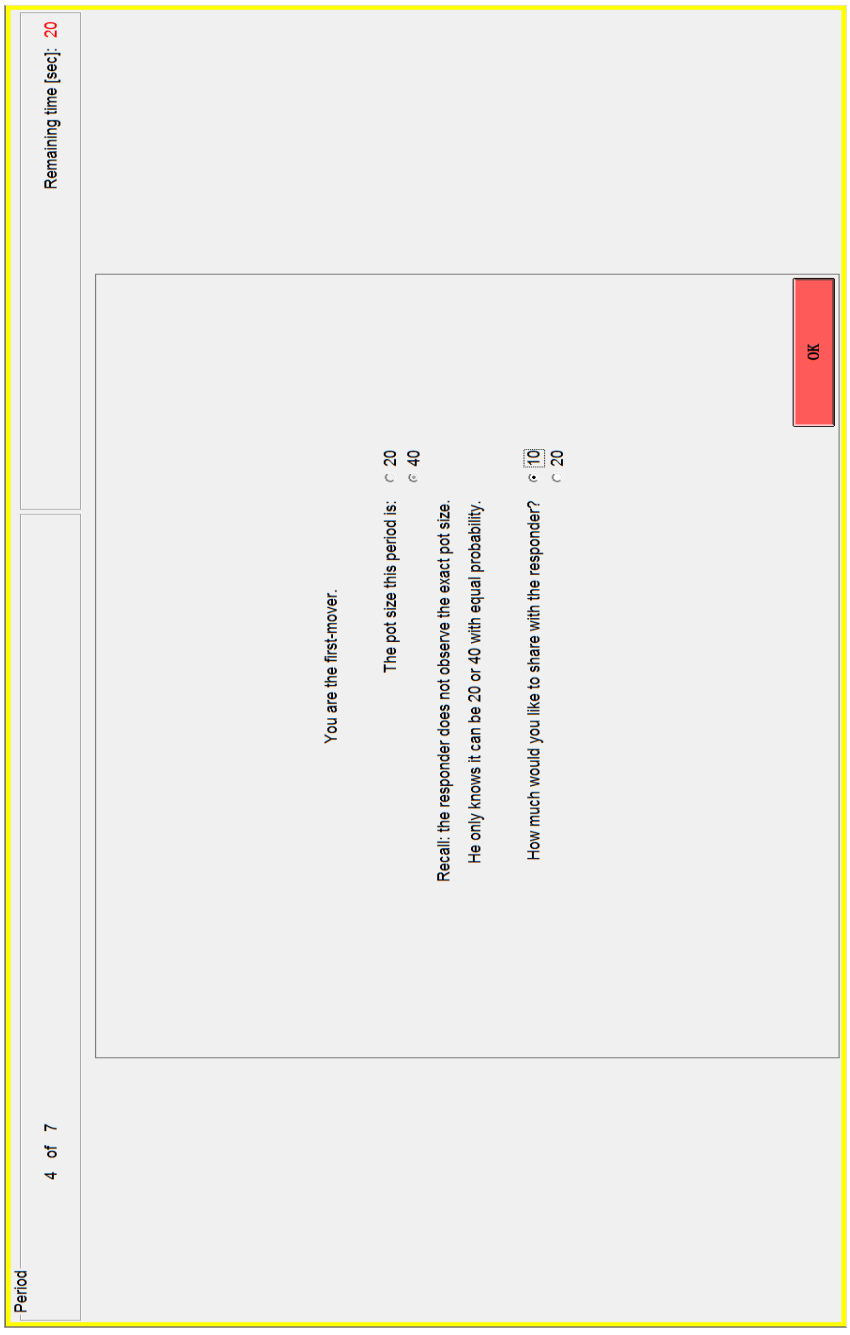


Figure 1.4: This screenshot displays what a responder might see in a given round. The responder decides in the case when his partner gives a low share (10 ECUs) how much he or she would like to reduce the partner's earnings. The table at the top shows how much the partner shared in three previous rounds and how much punishments were used.

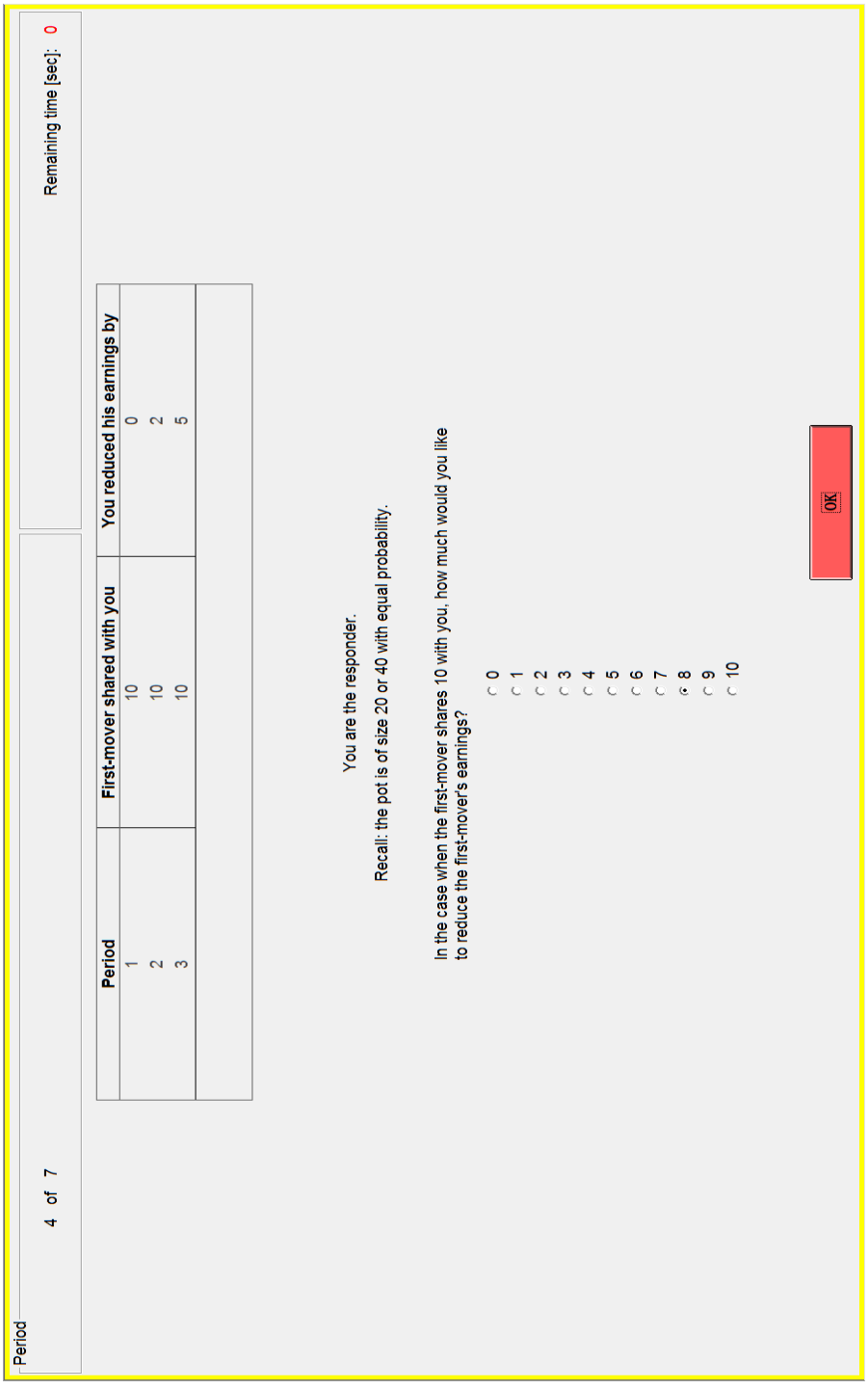


Figure 1.5: This screenshot displays a second screen a responder might see in a given round. The responder decides in the case when the partner gives a high share (20 ECUs) how much he or she would like to reduce the partner's earnings. The table at the top shows how much the partner shared in three previous rounds and how much punishments were used.

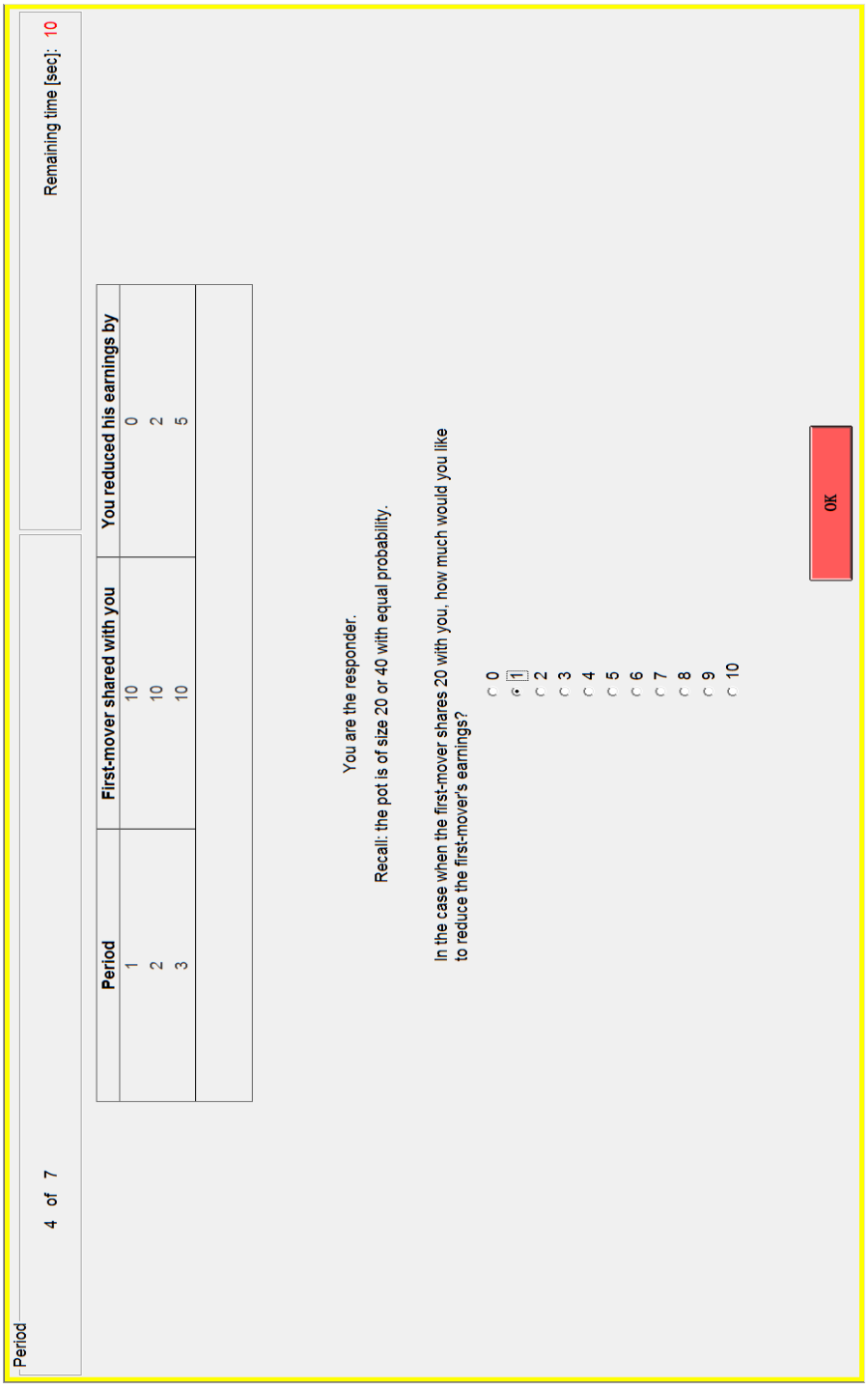


Figure 1.6: This screenshot displays a third screen a responder might see in a given round. The responder is shown whether his partner gives a low or high share. The responder then answers two questions on anger. The first one is 'How angry are you with the share your received?' The second one is 'How angry are you at the first-mover?' The table at the top shows how much the partner shared in previous rounds and how much punishments were used in return.

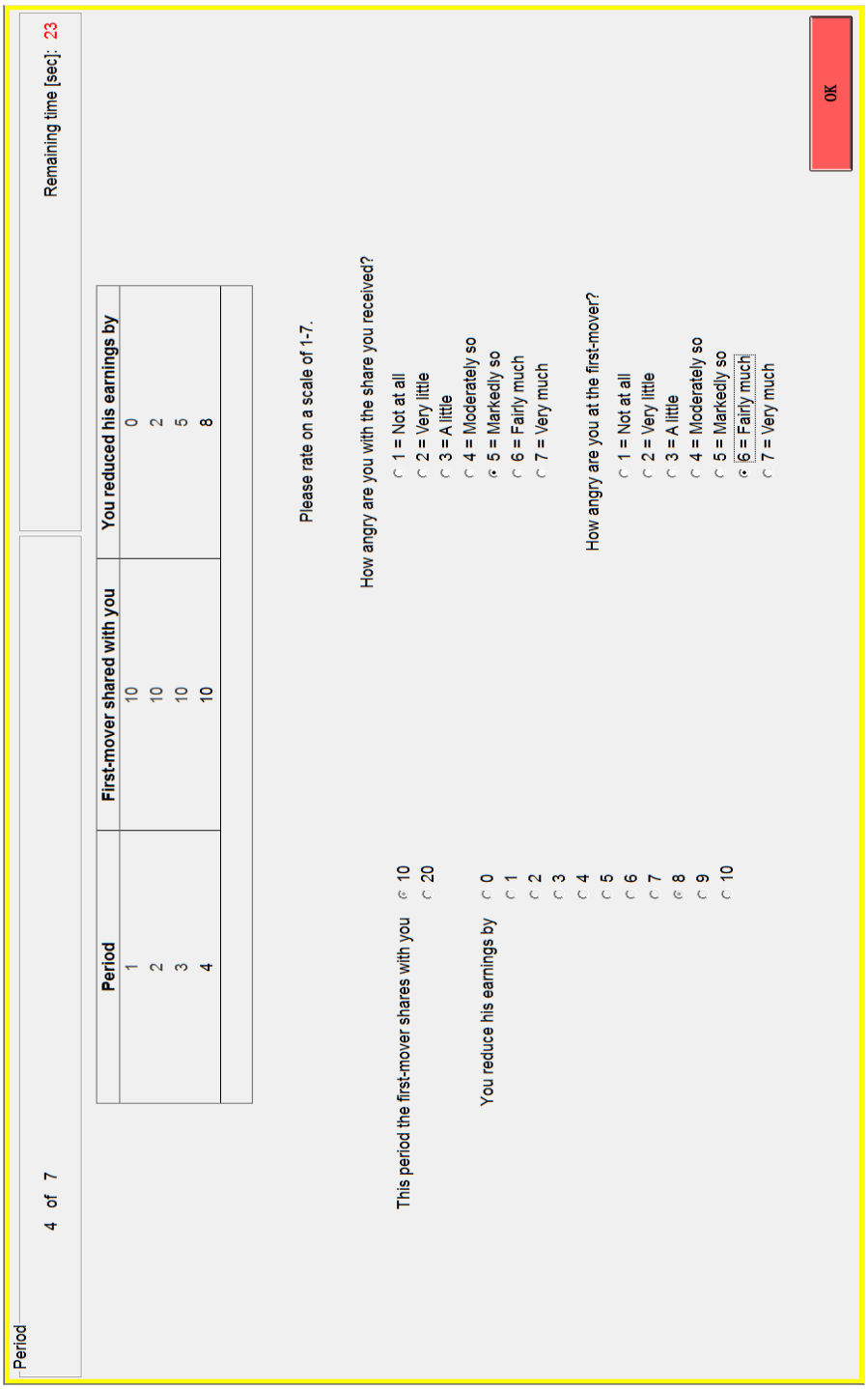


Figure 1.7: This screenshot displays a fourth screen a responder might see in round 4. The responder indicates two likelihood estimations. The first one is 'Given the share your received, what is your best estimate of the likelihood that the pot size this period is 40?' The second one is 'Assume the pot size is 40 next period. What is your best estimate of the likelihood that the first-mover will share 10 with you?' The table at the top shows how much the partner shared in four rounds and how much punishments were used in return.

Chapter 2

Intuition and Deliberation in Giving and Punishment

2.1 Introduction

Altruistic giving and punishing norm violators are important social and economic phenomena. In particular, evidence shows that the opportunity of punishment can increase social cooperation (Fehr & Gächter, 2000a). However, people tend to punish less when the cost of punishing rises, and individuals differ in the extent to which they respond to changes in the cost of punishing (Anderson & Putterman, 2006; Carpenter, 2007). Some punish despite having to pay a high cost of punishing, while others are quite sensitive to any changes in punishment cost. This study explores whether the way people process information can explain the individual differences in altruistic giving and punishment.

Prior research has studied the effect of response time in reaching a decision on the decision itself (e.g. Rand *et al.* , 2012; Sutter *et al.* , 2003). Some studies use the ultimatum game, in which a proposer makes an offer on how to split a sum of money with a responder, and the responder can then decide to accept or reject the offer. If the offer is accepted, both parties divide the money according to the proposed offer; if rejected, both parties get nothing. Rejecting a low offer is a way to punish the proposer although the responder also incurs a cost of not accepting the offer. Evidence has shown that the rate of punishment increases under time pressure (Sutter *et al.* , 2003) and decreases when decisions are delayed (Grimm & Mengel, 2011), suggesting

that punishing is driven by fast and intuitive thinking and reduced by slow and deliberate thinking. However, Rubinstein (2007) found that the response time of those who accepted a low offer was remarkably similar to cases where low offers were rejected. Thus, punishing of this kind is not necessarily correlated with quick responses. There is so far no conclusive evidence on the role of intuitive and deliberate thinking in punishment behaviour.

One controversial claim in the literature is that intuition leads to giving and cooperation. Rand *et al.* (2012) found in an experiment that subjects who reached their decisions more quickly were more cooperative. To explain this finding, Rand *et al.* (2014) proposed a theory that social norms shape our intuition. Since cooperation is usually advantageous, people form an intuition to cooperate in daily life and bring it to the experiment. However, several experiments using time pressure to induce intuitive responses have not replicated the same result in finding that cooperation is intuitive (Tinghög *et al.* , 2013; Verkoeijen & Bouwmeester, 2014). Furthermore, a more detailed analysis by Myrseth & Wollbrant (2015) showed that Rand *et al.* (2012) and Rand *et al.* (2014) misinterpreted their own data and there was no clear relationship between decision time and cooperation.

Instead of using the response time to measure thinking processes, the present study uses the well-established fact that some individuals are more intuitive than others and some are more deliberate than others (e.g. Epstein *et al.* , 1996). These individual differences can be reliably measured by the Rational-Experiential Inventory (REI), a self-report 40-item questionnaire, developed by Pacini & Epstein (1999). The broader theoretical basis of the REI comes from the dual-process literature in psychology (e.g. Kahneman, 2011; Evans & Stanovich, 2013), which distinguishes two separate systems of information-processing (intuitive and analytic) that jointly produce a variety of behaviours. The intuitive system is assumed to be impulsive, affective, automatic, fast, unconscious and effortless. The analytic system is assumed to be reflective, logical, deliberate, slow, conscious and effortful. More specifically, the REI is underpinned by Cognitive-Experiential Theory (Epstein, 2014), which has the unique feature of placing the intuitive and analytic systems in a theory of personality. Other independent studies have also confirmed that the REI produce both valid and reliable measures of individual differences (e.g. Witteman *et al.* , 2009; Hodgkinson *et al.* , 2009; Björklund & Bäckström, 2008).

The Rational-Experiential Inventory has two independent scales cor-

responding to the two systems. One scale called Need for Cognition (NC; adopted from Cacioppo & Petty, 1982) measures the extent to which an individual engages in and enjoys effortful analytic thinking. This scale contains 20 items, including 'I prefer complex problems to simple problems', 'I don't reason well under pressure', and 'I have no problem thinking things through carefully.' The NC was originally conceptualized to reflect the need to understand and make sense of the world (Cohen *et al.* , 1955). Since then, hundreds of studies have investigated how the NC correlates with various behaviours and personality variables (for a review Cacioppo *et al.* , 1996). For instance, people who report a high tendency to think deliberately think more about available options before making a decision (Levin *et al.* , 2000), have reduced sensitivity to framing effects and sunk costs (Carnevale *et al.* , 2011), and form attitudes by paying attention to issue-relevant information instead of peripheral cues (Petty & Cacioppo, 2012). In many respects, a higher NC does lead to more rational choices by eliminating behavioural biases, but not all biases (e.g. priming effect; Petty *et al.* , 2008).

The other scale, Faith in Intuition (FI), measures the extent to which an individual relies on intuitive feeling. It contains 20 items, such as, 'I believe in trusting in my hunches', 'if I were to rely on my gut feelings, I would often make mistakes', and 'intuition can be a very useful way to solve problems.' The FI scale was developed after the Need for Cognition scale in order to measure the intuitive dimension of the dual-process theory (Epstein *et al.* , 1996). Several studies have found that high FI scores are associated with increased reliance on heuristic rules in decision making (Shiloh *et al.* , 2002; Danziger *et al.* , 2006; Glaser & Walther, 2014). For instance, Alós-Ferrer & Hügelschäfer (2012) found that those who rely more on their intuition are more likely to fail to update their beliefs based on all available information because they overweight sample information. Mahoney *et al.* (2011) found that people with higher FI scores are more likely to show preference reversal due to different framing of the Asian Disease problem (Tversky & Kahneman, 1981).

This study experimentally examines the role of intuition and deliberation, as measured by Faith in Intuition and Need for Cognition, in social interactions when the cost of punishing varies. An online experiment is conducted with participants recruited via Amazon Mechanical Turk (MTurk). The experimental design consists of a two-player game, in which the first player can

decide how much to give to the second player out of a sum of money, and the second player can then decide how much to punish the first player at a cost. The game is repeated for two rounds with only the price of punishment changing between rounds. In the Price Decrease Treatment, the price of punishment is high in the first round and low in the second round. In the Price Increase Treatment, the price is low in the first round and high in the second round. These two treatments allow us to measure how responsive people are to changes in the price of punishment.

This experiment also allows us to compare the predictions from economic theories and dual-process theory of decision-making. The standard economic theory that assumes material self-interest predicts that no one will punish if punishing is costly and no one will give if no one would punish those who do not give. Recently developed theories that incorporate social preferences, such as inequality aversion (Fehr & Schmidt, 1999) and reciprocity (Dufwenberg & Kirchsteiger, 2004), are able to explain the observed willingness to give and punish. On the other hand, in the dual-process framework, the social heuristic hypothesis (Rand *et al.* , 2014) proposes that intuition is shaped by social norms and deliberation adjusts behaviour towards the personal optimum. Hence, we expect that people who tend to think more deliberately will give less and punish less. People who rely more on intuitive feelings will give more and punish more if there exists social norms that encourage generous giving and punishing selfishness.

We find that many participants are willing to give an equal share and many are willing to punish those who give nothing. But the price of punishment significantly reduces that number of punishers as well as the number of givers. Participants who tend to think more deliberately are less likely to punish, suggesting that deliberation might restrict the impulse to punish. The relationship between reliance on intuition and giving is positive but not statistically significant. We also find that high reliance on intuition is associated with a greater sensitivity of punishment to a price increase than to a price decrease, which might be explained by loss aversion (Kahneman & Tversky, 1979; Heidhues & Kőszegi, 2008).

This research contributes to the experimental literature that studies individual heterogeneity in altruistic preference (Andreoni & Miller, 2002) and in punishment behaviour (Anderson & Putterman, 2006; Carpenter, 2007). First, Andreoni & Miller (2002) show that people have varying degrees of altruism

and suggest that accounting for these differences is necessary for understanding choices. Our study advances this line of research by showing that individual tendency to give altruistically might be influenced by the reliance on intuition, rather than an aversion to think deliberately. Second, Anderson & Putterman (2006) demonstrate considerable, yet unexplained, individual differences in the propensity to punish and the sensitivity to punishment price. Our study reveals more by showing that the tendency to punish can be explained by an aversion to think deliberately and that the price sensitivity of punishment is affected by individual reliance on intuitive feelings.

2.2 Experimental Setup

2.2.1 Experimental Design

We design a two-player game which takes place over two rounds. In each round, the first player can decide how to split a sum of money ($4) between the participants, and the second player can then choose how much to punish the first player.

The price of punishment is defined as the amount player 2 must spend to remove one dollar from player 1. This price can be either high (50¢) or low (10¢). For instance, if player 2 wants to reduce player 1's earnings by two dollars and the price is 10¢, it will cost him 20¢. If the price is 50¢, the cost for removing two dollars will be $1. In the Price Decrease Treatment, subjects are assigned to play this game with the high price of punishing in the first round and then with the low price in the second round. To examine the effect of price order, we conduct the Price Increase Treatment which starts with low price first and high price second. In order to prevent player 2 from ending up with a negative payoff due to the cost of inflicting punishment, the second players are endowed another two dollars in each round.

To elicit punishment decisions, we use the 'strategy method'. This method requires player 2 to give a response for each feasible action of player 1 before player 2 is informed of player 1's actual choice.[1] Since only one action is selected by player 1, only in this specific case both players' decisions will affect payoffs. To simplify the game without losing the possibilities of selfish and

[1] Brandts & Charness (2011) survey the experimental literature and studied the impact of using the strategy method. The authors show that using strategy method can affect the level of punishment but not the qualitative results such as the effect of price on punishment.

equal divisions, we restrict player 1's choice set to three feasible actions: to give zero, one, or two dollars to player 2. This approach enables us to collect three contingent punishment decisions in response to player 1's three feasible actions, even though only one action is actually chosen by player 1.

With regards to the final payoffs, an equal division would be three dollars for each player if player 1 gives one dollar and player 2 refrains from punishing. Giving two dollars would then be more than fair, and giving zero dollars would be most selfish.

2.2.2 Online Survey and Participants

We conduct this experiment using an online survey with MTurk workers as participants. Since online participants fill out the survey independently without any temporal coordination with other participants, there is no possibility to give feedback about the partner's choices during the course of the experiment.[2] In practice, we first collect all survey responses from the participants and then match their responses for calculating final payoffs. All participants are told that their payoffs depend on their decisions as well as the decisions of others.

Our survey designates five pages for this section of the experiment. The first page includes instructions about the game and tells each participant that he or she is the second player[3] and that the price of punishment is 50¢ in the Price Decrease Treatment (or 10¢ in the Price Increase Treatment). To enhance the participant's comprehension of the game, we present three scenarios of possible plays of the game for the participant to compute the total cost of punishment and potential earnings. We find that 83.7% of the sample compute the answers correctly. On the second page, the participant is asked to indicate a punishment decision for each of the three feasible actions of player 1. On the third page, the participant is informed that he or she would interact with the same partner again in the same procedure as before, but the price of punishment is changed to 10¢ (or 50¢). The fourth page states that the participant is now the first player of the game and is matched with a different partner. The price of punishment is 50¢, and the participant has

[2] One limitation of online surveys is that only some types of experiments can be run (for a review, see Horton *et al.* , 2011).

[3] This is the starting point for all participants, and they also play the role of the first player later. This is possible because participants fill out the survey independently before their decisions are matched with the decisions of their partners.

to decide how to split \$4 with the partner. On the fifth page, the price of punishment is now 10¢, and the participant has to split another \$4 with the partner. These five pages together independently elicit all of the contingent decisions from each participant.

Note that not receiving feedback on their partner's choices during the course of the experiment removes some contaminants that would normally be generated in repeated interactions. Thus, the decisions made in the second round would mostly reflect the new price of punishment, rather than the interaction in the first round. However, there might still be some effects of taking the first price of punishing as a reference point for second round decisions.

325 participants were recruited via MTurk in five sessions in September 2014. Participants spent 12 minutes on average to complete the survey and were paid a fixed rate of \$1.50 for completion (the hourly rate for MTurk workers was \$6). All the participants were told at the beginning of the survey that only one participant in each session would be randomly selected to receive an additional payment based on the choices in the experiment. The additional payment included payoffs as player 1 and player 2. A total of five participants were selected to receive additional payments, which on average were \$11.53. The entire survey was created by using Qualtrics survey software (Qualtrics, Proto, UT).

Although online experiments usually use a lower level of financial incentives than physical laboratory experiments, there are reasons to believe that this does not affect data quality. First, MTurk workers are willing to accept lower wages because they do not have to pay travel costs and they have flexibility to choose when to work. Second, Mason & Watts (2010) found that when wages increased from \$0.01 to \$0.10 per task, MTurk workers completed a larger number of tasks, but there was no difference in the quality of work. This finding is consistent with the modal result of a survey of 74 experiments done by Camerer & Hogarth (1999), who showed that financial incentives have no effect on mean performance in most experiments, except in tasks that are responsive to better effort. Lastly, Horton *et al.* (2011) used MTurk to replicate the results of three classic experiments: the level of cooperation in Prisoner's Dilemma game, the framing effect demonstrated by Tversky & Kahneman (1981), and the priming effect on cooperation.

2.3 Predictions

This section first describes the predictions of economic theories for our experiment. We then consider other hypotheses based on the dual-process theory of decision making.

The standard economic theory assumes that players have common knowledge that every player is rational in maximizing their own material payoffs. For the experiment, the prediction of the subgame perfect equilibrium is the following: Player 2 will not punish, regardless of player 1's giving decision, because punishing is costly. Player 1 will not give anything to player 2 because he knows that player 2 will not punish him even if he gives nothing.

Yet, a substantial number of people in this type of experiments do not follow the prediction of rationality or material self-interest. To explain this behaviour, recently developed theories incorporate social preferences such as inequality aversion and reciprocity. Models by Fehr & Schmidt (1999) and Bolton & Ockenfels (2000) assume that people dislike inequality and aim to reduce the payoff difference between players. Dufwenberg & Kirchsteiger (2004) and Falk & Fischbacher (2006) incorporate a notion of reciprocity that kind or unkind intentions behind an action can trigger reciprocal behaviour. All these models predict that for sufficiently strong concern about equality or reciprocity, player 1 will give a positive amount of money to player 2, and player 2 will punish player 1 who gives an amount that is considered unequal or unkind.

The prediction on the impact of changes in the price of punishment on punishment decisions is less clear. For instance, Fehr & Schmidt's model of inequality aversion can generate multiple equilibria. For a sufficiently strong aversion to inequality, player 2 will punish more when the punishment price is high. The reason is that the amount of payoff difference between two players that is reduced by a punishment unit (\$1) is smaller at a high price ($\$1 - 50¢ = 50¢$) than at low price ($\$1 - 10¢ = 90¢$). Hence, more punishment units are required at a high price to reduce the same amount of payoff difference than at low price. But there are also plausible equilibria where player 2 will punish less when the punishment price is high.

Our main interest is to look at the relation between the predictions of economic theories and dual-process theory of decision making. Rand *et al.* (2014) proposed a social heuristic hypothesis that intuition is shaped by social

norms and deliberation adjusts behaviour towards utility maximization. This hypothesis implies that individuals with a higher Need for Cognition are more likely to follow the rationality predictions, which are to refrain from giving and punishing. If the everyday norms that shape our intuition are to give generously and to punish selfishness, we would expect that people with a higher Faith in Intuition are more likely to have social preferences that motivate giving and punishing.

2.4 Results

Table 2.1 separately presents a summary of the experimental data statistics and the characteristics of MTurk workers for Price Decrease Treatment ($N = 161$) and Price Increase Treatment ($N = 164$). More than 65% of the participants are white and an almost equal number of male and female. About 40% of the participants are younger than twenty-nine years old, and 10% are at least fifty years old. More than half had attained at least a Bachelor's degree. The last column of Table 2.1 shows that individual characteristics, such as gender, age, race, education, etc., are statistically identical across treatments at 5% level. This suggests that participants are well randomized across treatment conditions.

Table 2.1 also checks whether there is a treatment effect on the participant decisions making. For instance, when the price of punishing is high, the distribution of punishment in the Price Decrease Treatment is not significantly different from the Price Increase Treatment. Similarly, no treatment effect is found when the price of punishing is low. Thus, the second punishment decision is not biased on average by the exposure to the first round. However, this does not preclude the possibility that a price increase will affect the price sensitivity of punishment in a different way than a price decrease.

Faith in Intuition varies in our sample from 1.05 to 5.00 with an average of 3.32. Need for Cognition varies from 1.65 to 5.00 with an average of 3.70. The correlation between Need for Cognition and Faith in Intuition is close to zero and insignificant (Spearman rank correlation: -0.075, $p = 0.179$). A scatterplot of these two scales also shows no clear relationship between intuitive and deliberate thinking (see Figure 2.9 in the Appendix). This is consistent with the previous finding that these two scales are independent dimensions of the dual-process model (Pacini & Epstein, 1999). We also find

Table 2.1: Descriptive Statistics by Treatment

	Treatment				
	Price Decrease		Price Increase		
Variables	**Mean**	**S. D.**	**Mean**	**S. D.**	***p*-value**
High price of punishment (50¢)	*1st round*		*2nd round*		
Punishment ($): if partner gives $0	$1.22	1.40	$1.23	1.48	0.69
if partner gives $1	$0.75	1.00	$0.79	1.06	0.99
if partner gives $2	$0.36	0.86	$0.38	0.86	0.76
Punishing frequency: if partner gives $0	0.58	0.49	0.53	0.50	0.37
if partner gives $1	0.52	0.50	0.49	0.50	0.66
if partner gives $2	0.22	0.41	0.23	0.42	0.79
Low price of punishment (10¢)	*2nd round*		*1st round*		
Punishment ($): if partner gives $0	$1.90	1.62	$2.06	1.64	0.36
if partner gives $1	$1.17	1.22	$1.15	1.17	0.97
if partner gives $2	$0.45	1.01	$0.42	0.79	0.34
Punishing frequency: if partner gives $0	0.69	0.46	0.73	0.44	0.46
if partner gives $1	0.62	0.49	0.66	0.48	0.49
if partner gives $2	0.23	0.42	0.29	0.46	0.21
Giving when punishment price is high	$1.38	0.72	$1.38	0.75	0.82
Giving when punishment price is low	$1.47	0.73	$1.46	0.72	0.84
Faith in Intuition (FI)	3.31	0.71	3.33	0.71	0.50
Need for Cognition (NC)	3.76	0.69	3.64	0.65	0.08
Time used to complete survey (minutes)	12.80	6.72	12.15	6.65	0.40
Female	0.52	0.50	0.48	0.50	0.51
White	0.67	0.47	0.65	0.48	0.82
Bachelor's degree	0.52	0.50	0.52	0.50	1.00
Cohabiting	0.56	0.50	0.46	0.50	0.08
Unemployed	0.18	0.39	0.18	0.39	1.00
Household income below $25,000	0.29	0.46	0.38	0.49	0.10
Age from 18 to 28 year olds [a]	0.39	0.49	0.40	0.49	1.00
Age from 29 to 34 year olds	0.29	0.45	0.26	0.44	0.71
Age from 35 to 49 year olds	0.21	0.41	0.24	0.43	0.60
Age at least 50 years old	0.11	0.32	0.10	0.31	0.86
Number of subjects	161		164		

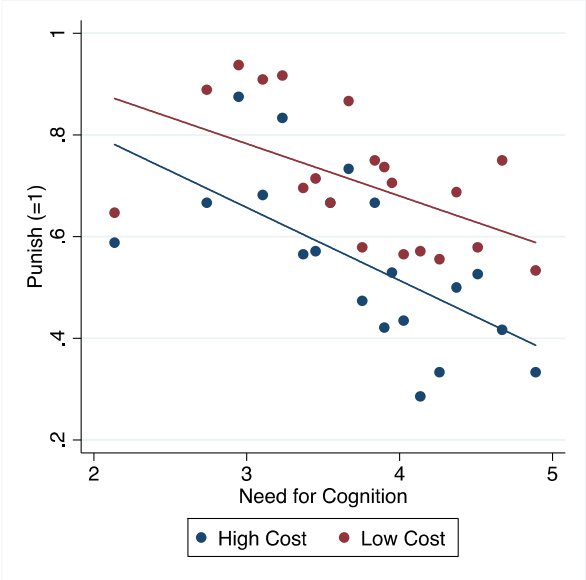
Notes: This table describes the experimental data and the characteristics of MTurk workers by treatment. The first four columns show the treatment averages and standard deviations, and the last column reports the p-values (Fisher's exact tests for binary variables and Mann-Whitney tests for non-binary variables) for the null hypothesis of identical distributions between treatments.

[a] A few subjects in this age group might be 29 years old because the cut-off in one session was 29 instead of 28.

that people high in Need for Cognition tend to take longer time to complete the survey (Spearman rank correlation: 0.115, $p = 0.039$; see Figure 2.10). This suggests that the self-report measure of deliberate thinking does have relevance to actual behaviour in our experiment.

In the case of player 1 giving zero dollars, 71.1% of subjects choose to punish when the cost of punishment is low. When the cost is high, the proportion of punishers drops significantly to 55.7% (Matched-pairs signed-rank test, $z = 6.934$, $p < 0.001$). If deliberation increases the awareness of the cost of punishing, we would expect that subjects who are higher in Need for Cognition are less likely to punish. Figure 2.1 portrays this result. The link between deliberation and cost-sensitivity is further demonstrated by looking at a subject's pair of punishment decisions at high and low costs. About 55.4% of subjects choose to punish at both prices, and they tend to score less on Need for Cognition than the 28.6% who do not punish at either price (Mann-Whitney test, $z = 3.776$, $p < 0.001$). There is a third group of subjects, 15.7%, who switch from punishing at low price to not punishing at high price. These price-sensitive punishers also tend to be more deliberate than the first group of consistent punishers (Mann-Whitney test, $z = 2.479$, $p = 0.013$).

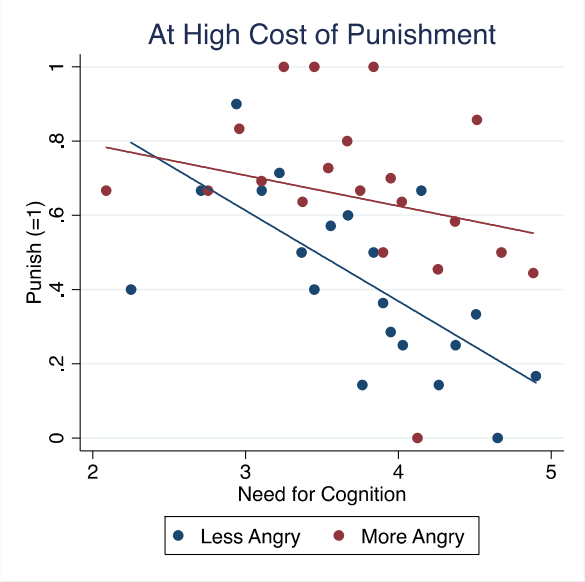
Figure 2.1: Deliberate People Are Less Likely to Punish



Notes: This binned scatter plot shows that relationship between the decision to punish and Need for Cognition. Subjects are grouped into twenty equal-sized bins based on their Need for Cognition scores, and each bin indicates its average Need for Cognition and its proportion of subjects who punish.

According to the dual-process theory, what comes to mind intuitively is sometimes in conflict with what derives from deliberate thought. The outcome behaviour depends on situational factors and individual traits. Subjects in this experiment are likely to face a similar conflict in making the decision to punish. Angry feelings or concerns for fairness can create a desire to punish, while the cost of punishing might dampen this desire. Thus, deliberate people might find it more difficult to override their urge to punish when they are angry. To test this hypothesis, we divide subjects into two groups based on their reported level of anger. When receiving zero dollars, about 55.4% of subjects report that their anger level is at least five out of seven. Figure 2.2 shows that deliberate people tend to give in to punishing when they report higher levels of anger.

Figure 2.2: Deliberate People Give in to Punishing When Angry



Notes: This binned scatter plot shows that relationship between the decision to punish and Need for Cognition when the cost of punishment is high. Subjects are grouped into twenty equal-sized bins based on their Need for Cognition scores, and each bin indicates its average Need for Cognition and its proportion of subjects who punish. More Angry refers to subjects who report a level of anger that is at least five out of seven.

2.4.1 Determinants of Giving

Table 2.2 shows the distribution of subjects based on the two giving decisions when player 2 faces high and low prices of punishment. Almost half of the

sample give two dollars in both decisions. Only 19.1% give one dollar for both decisions and 10.8% give zero dollars. This suggests that most people did not take into account the two dollars endowed to player 2. We also find that subjects who give two dollars for both decisions tend to score higher on Faith in Intuition than those who give only one dollar (Mann-Whitney test, $z = 2.882$, $p = 0.004$). However, the intuitive response is not simply giving more because subjects who give nothing are also more intuitive than those who give one dollar at 10% significance level (Mann-Whitney test, $z = 1.762$, $p = 0.078$). Therefore, the intuitive response for different individuals could be either giving two or zero dollars, but not one dollar. Table 2.2 also shows that 15.1% of subjects give less when player 2 faces a high cost of punishment, but they are particularly intuitive or deliberate. Less than 10% of subjects give more when the cost of punishment high. These subjects might be confused because more than a third of them fail to answer comprehension questions correctly.

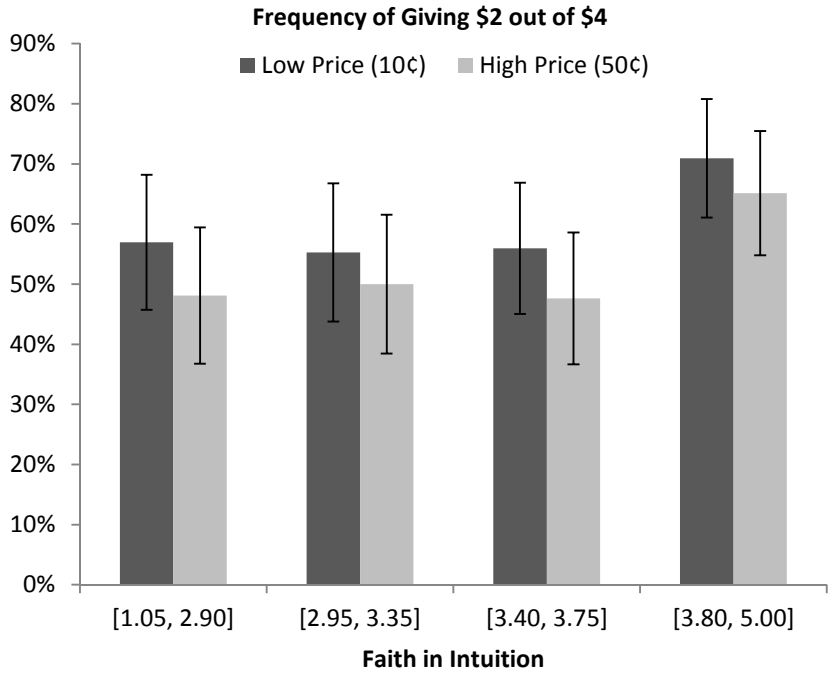
Table 2.2: Distribution of Giving Decisions (%)

		High Cost of Punishment (50¢)		
		Give $2	Give $1	Give $0
Low Cost of Punishment (10¢)	Give $2	47.1	11.1	1.9
	Give $1	4.9	19.1	2.2
	Give $0	0.9	2.2	10.8

Notes: Each subject makes two giving decisions: one when thier partner faces a low cost of punishment, and the other a high cost of punishment. Each cell indicates the percentage of subjects who make a particular pair of giving decisions. For example, 47.1% of subjects choose to give $2 under both conditions.

Figure 2.3 illustrates the relationship between Faith in Intuition and the frequency of subjects giving an equal share of two dollars. Subjects are divided based on Faith in Intuition into four groups of approximately equal size. The three lowest intuitive groups have similar frequencies of giving an equal share, ranging from 48% at high price of punishing to 57% at low price of punishing. However, the most intuitive group ($N = 86$) give an equal share of two dollars about 17.5 percentage points more frequently than the second most intuitive group ($N = 84$) when the second player's punishment price is high (Fisher's exact test, $p = 0.030$) and about 15 percentage points more frequently when the price is low (Fisher's exact test, $p = 0.056$).

Figure 2.3: Faith in Intuition and Giving Frequency.



Notes: Subjects are divided into four groups based on the Faith in Intuition scores. The number of subjects from the lowest to the highest group: N = 79, 76, 84, 86. For each group the first bar is the giving frequency when player 2's price of punishing is low and the second bar is when the price of punishing is high. Each error bar is twice the standard error of the mean, approximating 95% confidence interval.

More formally, in column 1 of Table 2.3 we use ordered logit model to estimate the first player's decision to give. The estimates are the marginal effects in predicting the choice of giving two dollars at the means of regressors. These marginal effects predict the tendency to give an equal share, not just to give more, because ordered logit model differentiates between an increase in giving from \$0 to \$1 and an increase in giving from \$1 to \$2. For comparison, column 2 of Table 2.3 reports OLS estimates.

The independent variables in Table 2.3 include *High punishment price*, ln(*Faith in Intuition*), ln(*Need for Cognition*), *Price Increase Treatment* (1 = Yes), and individual characteristics. In column 1, the first coefficient on the exogenous dummy variable *High punishment price* is −0.064 and significant ($p = 0.005$), indicating that the first players are 6.4 percentage points less likely to give an equal share if they anticipate the second player faces the high price of punishing (50¢). This corresponds to earlier results that people tend

Table 2.3: Determinants of the First Player's Decision to Give

Dependent Variable: Giving $\{0, 1, 2\}$	Ordered Logit	OLS
Independent Variables	(1)	(2)
High punishment price	-0.064***	-0.080**
	(0.023)	(0.031)
ln(*Faith in Intuition*)	0.110	0.070
	(0.103)	(0.144)
ln(*Need for Cognition*)	-0.112	-0.142
	(0.124)	(0.159)
Price Increase Treatment	-0.005	-0.001
	(0.051)	(0.072)
Female	-0.009	-0.019
	(0.053)	(0.076)
White	0.015	0.030
	(0.057)	(0.084)
Bachelor's degree	-0.240***	-0.350***
	(0.051)	(0.077)
Cohabiting	-0.063	-0.083
	(0.056)	(0.080)
Unemployed	0.112	0.122
	(0.069)	(0.093)
Household income < $25,000	-0.084	-0.101
	(0.054)	(0.077)
Age from 29 to 34	0.068	0.085
	(0.059)	(0.088)
Age from 35 to 49	0.026	0.023
	(0.073)	(0.105)
Age at least 50	0.186**	0.246**
	(0.079)	(0.122)
Constant		1.734***
		(0.294)
Subjects	325	325
Observations	650	650
R^2		0.093

Notes: The dependent variable is how much player 1 gives to player 2, which can be \$0, \$1 or \$2. Estimates in column 1 are the marginal effects in predicting the choice of giving \$2 in ordered logit model. OLS estimates in column 2. The binary variable *High punishment price* indicates whether the price of punishing is 50¢ (versus 10¢). The variable ln(*Need for Cognition*) is the logarithm of the degree to which subjects enjoy deliberate thinking. The variable ln(*Faith in Intuition*) is the logarithm of the degree to which subjects rely on their intuitive feeling. Standard errors in parentheses are clustered at the subject level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

to give less when they anticipate that others will be less prone to punish in response to a higher price of punishment. The variable ln(*Faith in Intuition*) in the second row of Table 2.3 is the natural logarithm of the Faith in Intuition score. Its point estimate in column 1 is positive but insignificant ($p = 0.282$). In the third row, the variable ln(*Need for Cognition*) is the natural logarithm of the Need for Cognition score. Its point estimate is negative but insignificant at 10% level, suggesting that there is no clear evidence that greater engagement in effortful thinking reduces the likelihood of giving an equal share of two dollars.

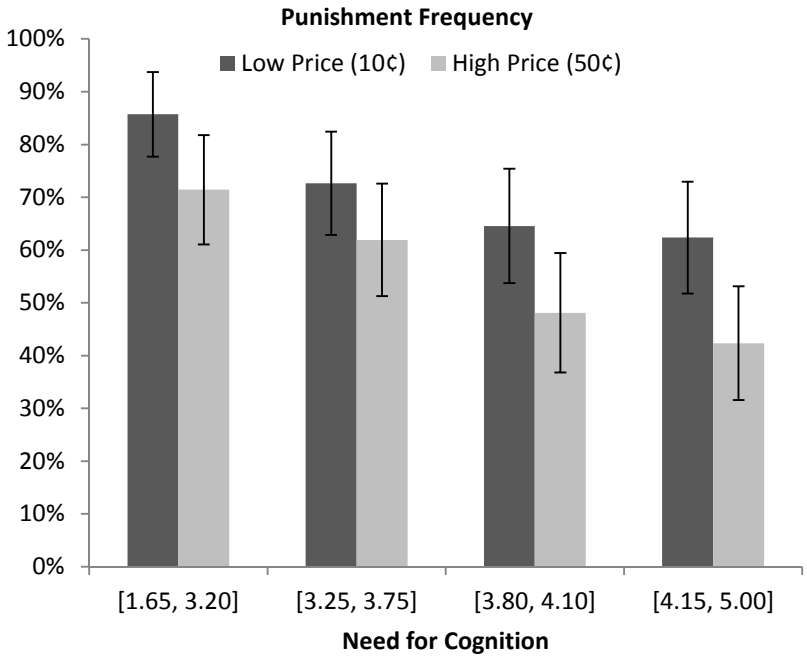
2.4.2 Determinants of Punishment

In each round, player 2 made three punishment decisions contingent on the three possible actions of player 1. The reasons for punishment are likely to differ in nature and vary in degree across these three cases. In the case when player 1 gives an equal share of two dollars, punishing is likely to be driven by spite, rather than fairness. In the case when player 1 gives less (one or zero dollars), people are more likely to punish due to social preferences, such as reducing inequality or reciprocating unkind behaviour. Thus, we analyse case by case the punishment decisions and examine the role of intuition and deliberation in the motivations behind punishment.

Figure 2.4 illustrates a negative relationship between the frequency of punishing and the tendency to think deliberately in the case when player 1 gives nothing. Based on the Need for Cognition scores, subjects are divided into four groups of approximately equal size. In the least deliberate group ($N = 77$), the proportion of punishers is 86% when the price of punishing is low. The proportion of punishers remains as high as 71% when the price of punishing is high. In comparison, the most deliberate group ($N = 85$) choose to punish about 29 percentage points less frequently when the punishment price is high (Fisher's exact test, $p < 0.001$) and about 23 percentage points less frequently when the punishment price is low (Fisher's exact test, $p = 0.001$). Thus, subjects who tend to think more deliberately are associated with lower likelihood of punishing.

Figure 2.5a shows that reliance on intuition does not have a clear linear relationship with the frequency to punish in the Price Decrease Treatment, but the Faith in Intuition has a surprising relationship with the responsiveness to

Figure 2.4: Need for Cognition and Punishment Frequency

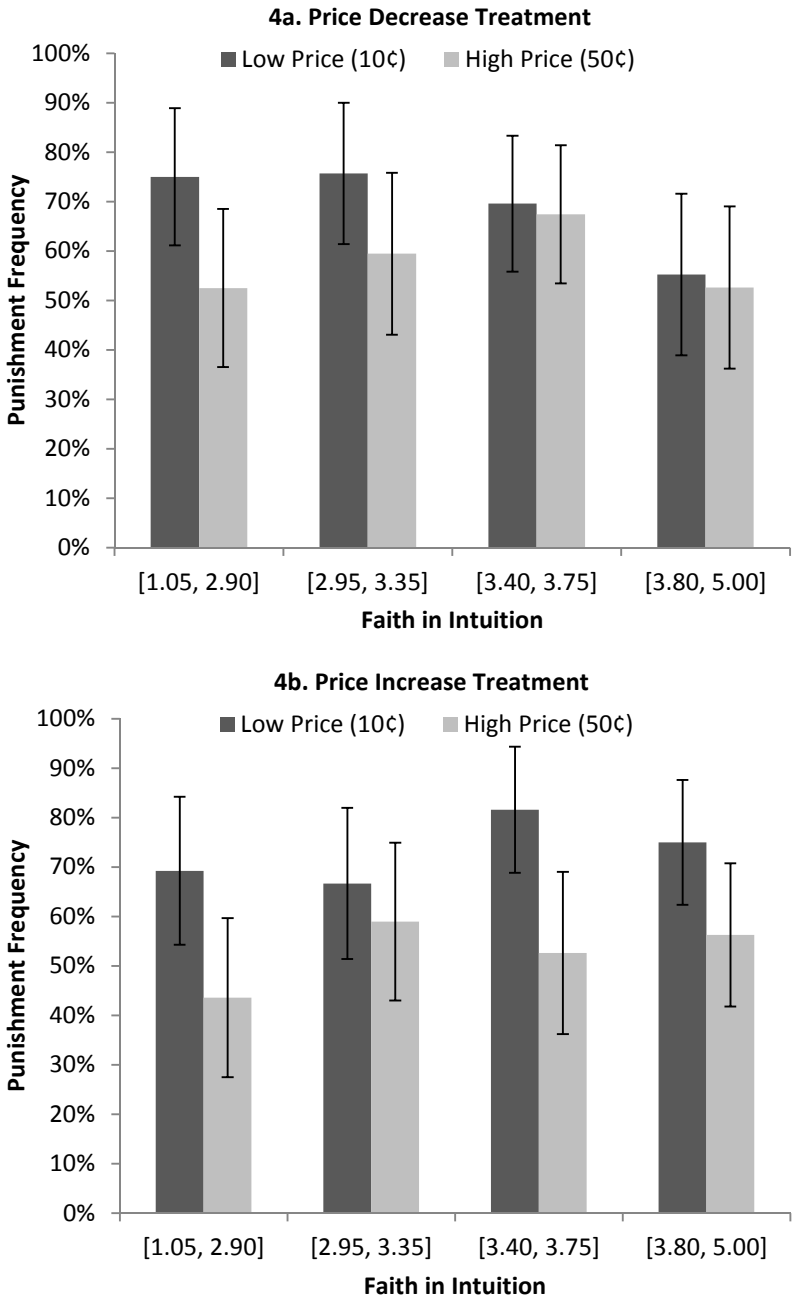


Notes: Subjects are divided into four groups based on the Need for Cognition scores. The number of subjects from the lowest to the highest group: $N = 77, 84, 79, 85$. For each group the first bar is the punishment frequency when the price of punishing is low and the second bar is when the price of punishing is high. Punishment frequencies reported are contingent on the case when player 1 gives nothing, as social preferences are more likely to be responsible for punishment. Each error bar is twice the standard error of the mean, approximating 95% confidence interval.

changes in the price of punishing. In the least intuitive group ($N = 40$), the proportion of subjects who switch from not punishing at a high price to punishing at low price is about 23% (as measured by the difference between the two bars of punishment frequencies at low and high prices for the group). In comparison, only about 3% of the most intuitive group ($N = 38$) are responsive in switching from not punishing at a high price to punishing at low price. Moreover, the difference in this proportion of price responsive subjects between the most and the least intuitive groups is large and significant (Fisher's exact test, $p = 0.014$). Thus, the most intuitive participants are less responsive to a decrease in the price of punishment than the least intuitive subjects.

Figure 2.5b illustrates a slightly different relationship between the Faith in Intuition and punishment frequency for the Price Increase Treatment. In the most intuitive group ($N = 48$), the proportion of subjects who switch from

Figure 2.5: Faith in Intuition and Punishment Frequency by Treatment



Notes: Subjects are divided into four groups based on the Faith in Intuition scores. For each group the first bar is the punishment frequency when the price of punishing is low and the second bar is when the price of punishing is high. Each error bar is twice the standard error of the mean, approximating 95% confidence interval. **4a.** In the Price Decrease Treatment, the number of subjects from the lowest to the highest intuitive group: $N = 40, 37, 46, 38$. **4b.** In the Price Increase Treatment, the number of subjects from the lowest to the highest intuitive group: $N = 39, 39, 38, 48$.

punishing at a low price to not punishing at a high price is almost 19% (as measured by the difference between the two bars of punishment frequencies at low and high prices for the group). If we compare this proportion with that of the most intuitive group in Figure 2.5a, we can see that the decision to punish is less responsive to a price decrease than to a price increase (Fisher's exact test, $p = 0.038$). On the other hand, the least intuitive group are similarly responsive to a price increase and to a price decrease (Fisher's exact test, $p = 0.797$). This suggests that only people high in Faith in Intuition display differential price sensitivity for punishment, depending on the exposure to the price of punishment in the first round.

Table 2.4 presents regression analyses on the second player's decision to punish contingent on the three cases of player 1's actions. Column 1 includes the punishment decisions in the case when player 1 gives \$0. Estimates reported are the marginal effects on punishing at the means of regressors in a logit model. The independent variables in column 1 of Table 2.4 include *High punishment price*, ln(*Need for Cognition*), ln(*Faith in Intuition*), *Price Increase Treatment* (1=Yes), and individual characteristics. The negative and significant estimate on *High punishment price* means that the high price of punishment (50¢) as opposed to low price (10¢) causes the likelihood of punishing to decrease by 16.1 percentage points ($p < 0.001$). This number measures how likely subjects are to switch the decision to punish due to price changes, and thus it represents the price sensitivity of punishment. The significantly negative estimate on ln(*Need for Cognition*) ($p = 0.021$) means that an increase in Need for Cognition score from 3 to 4 (on a scale ranging from one to five) decreases the likelihood of punishing by 10.9 percentage points. This confirms the negative relationship seen in Figure 2.4 between punishment frequency and the tendency to think deliberately.

We also find significant effects of punishment price and deliberate thinking on the likelihood to punish in column 2 of Table 2.4, which includes the punishment decisions in the case when player 1 gives \$1. In column 3 of Table 2.4, the punishment decisions are made in the case when the partner gives \$2. Even though these punishments can be considered spiteful, a high punishment price still decreases the likelihood of punishing. Therefore, subjects who tend to think more deliberately are less likely to punish, regardless of the cases when player 1 gives more or less.

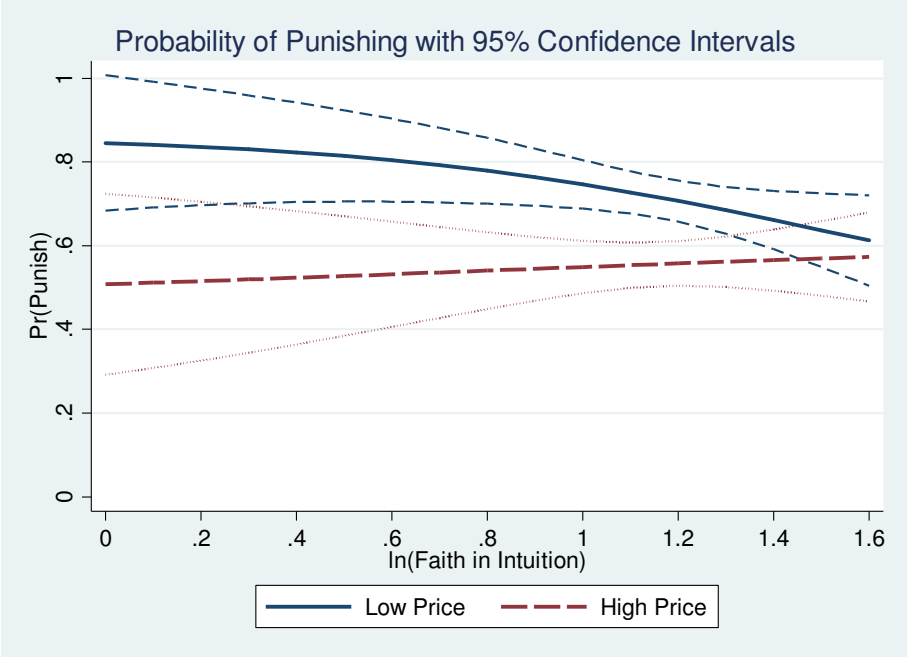
Although the coefficient on ln(*Faith in Intuition*) in column 1 of Table

Table 2.4: Determinants of the Second Player's Decision to Punish

Dependent Variable:	Logit Model		
Punish $\in \{0, 1\}$	If partner	If partner	If partner
Independent Variables	gives \$0	gives \$1	gives \$2
High punishment price	-0.161***	-0.143***	-0.038**
	(0.022)	(0.021)	(0.016)
ln(*Need for Cognition*)	-0.390**	-0.554***	-0.244**
	(0.168)	(0.162)	(0.097)
ln(*Faith in Intuition*)	-0.070	0.074	0.137
	(0.099)	(0.098)	(0.084)
Price Increase Treatment	-0.020	-0.019	0.029
	(0.051)	(0.053)	(0.043)
Female	-0.013	-0.034	-0.048
	(0.053)	(0.056)	(0.045)
White	-0.065	-0.140**	-0.125**
	(0.061)	(0.060)	(0.051)
Bachelor's degree	-0.033	-0.031	0.037
	(0.055)	(0.058)	(0.048)
Cohabiting	0.036	0.032	0.068
	(0.057)	(0.060)	(0.046)
Unemployed	0.063	0.017	-0.060
	(0.066)	(0.070)	(0.060)
Household income < \$25,000	0.067	0.069	0.114**
	(0.056)	(0.058)	(0.050)
Age from 29 to 34	-0.010	-0.018	-0.073
	(0.067)	(0.066)	(0.046)
Age from 35 to 49	-0.152**	-0.098	-0.116**
	(0.076)	(0.078)	(0.047)
Age at least 50	-0.103	-0.041	-0.143***
	(0.099)	(0.099)	(0.051)
Subjects	325	325	325
Observations	650	650	650

Notes: The dependent variable is the decision to punish or not. Estimates are the marginal effects on punishing at the means of regressors in logit model. The dummy variable *high punishment price* equals one when the price of punishing is 50¢ (versus low price of 10¢). The variable ln(*Need for Cognition*) is the natural logarithm of the degree to which subjects enjoy effortful analytic thinking on a scale from one to five. The variable ln(*Faith in Intuition*) is the logarithm of the degree to which subjects rely on intuitive feeling on a scale from one to five. The dummy variable *Price Increase Treatment* takes the Price Decrease treatment as the default. Robust standard errors in parentheses are clustered at the subject level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Figure 2.6: Predictive Probability on the Decision to Punish by Faith in Intuition

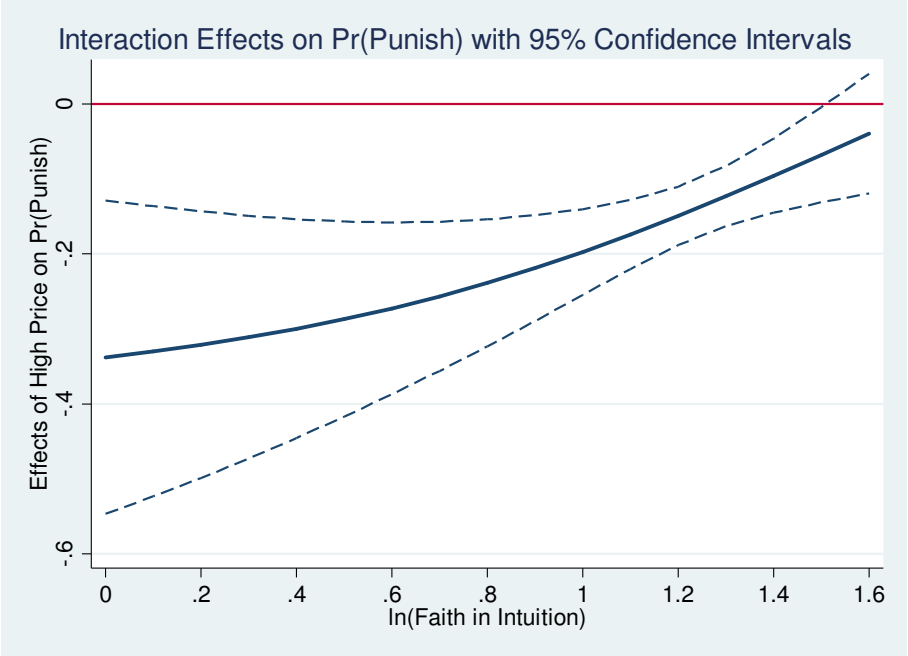


Notes: At high and low prices of punishment, the confidence intervals for the probabilities of punishing overlap as Faith in Intuition score reaches the higher end. All interaction terms between Faith in Intuition, punishment price and treatment dummy are added to the logit regression in column 1 of Table 2.4.

2.4 is insignificantly different from zero, we find significant interaction effects between individual reliance on intuition and the price of punishment. Figure 2.6 shows the predictive probability of punishing at a given price of punishment over ln(*Faith in Intuition*). When the punishment price is low, the predictive probability of punishing ranges from 85% for the least intuitive to 61% for the most intuitive. The probability at a high price ranges from 51% for the least intuitive to 57% for the most intuitive. At high and low prices of punishment, the confidence intervals for the predictive probabilities begin to increasingly overlap as the Faith in Intuition score reaches the higher end. This suggests that subjects with greater reliance on intuition are more likely to have the same likelihood of punishing at high and low prices.

Moreover, Figure 2.7 shows the average marginal effects of high punishment price on the probability of punishing over ln(*Faith in Intuition*). This effect measures the price sensitivity of punishment and is negative on average at all levels of Faith in Intuition. However, the negative price sensitivity of

Figure 2.7: Price Sensitivity of the Decision to Punish by Faith in Intuition



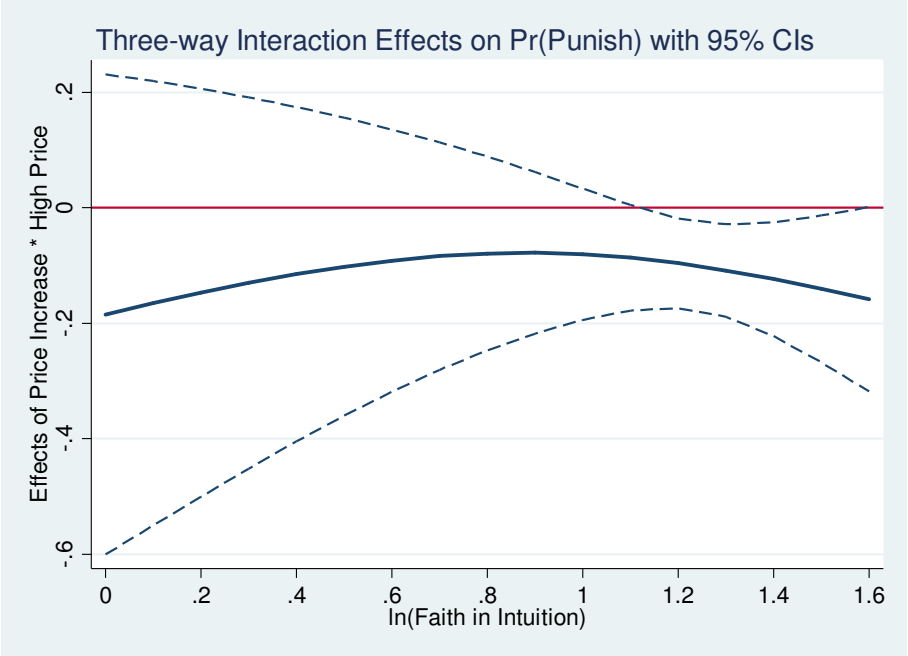
Notes: The negative price sensitivity of punishment becomes smaller in magnitude for high Faith in Intuition scores. All interaction terms between Faith in Intuition, punishment price and treatment dummy are added to the logit regression in column 1 of Table 2.4.

punishment becomes smaller in magnitude for subjects who tend to rely more on their intuitive feelings. For example, a subject who scores 3 on Faith in Intuition is 17.5 percentage points less likely to punish if the punishment price is high, while a subject who scores 4 on Faith in Intuition is only 9.9 percentage points less likely to punish if the punishment price is high. Thus, more intuitive participants are less sensitive to changes in the price of punishment than the less intuitive participants.

Finally, we find that subjects with high reliance on intuitive feelings display heterogeneous price sensitivity of punishment. Figure 2.8 shows three-way interaction effects between Price Increase Treatment (versus Price Decrease), high punishment price (versus low price) and ln(*Faith in Intuition*). The effect of Price Increase Treatment on the price sensitivity of punishment is negative on average at any level of Faith in Intuition, indicating that decisions to punish are more responsive to price changes in Price Increase Treatment than in Price Decrease Treatment. However, this additional sensitivity to a price increase is not significant at 5% level until the Faith in Intuition score rises above 3.3.

This suggests that only subjects with high reliance on intuition respond to a price increase more than to a price decrease. This asymmetric price sensitivity might be a result of loss aversion.

Figure 2.8: Effects of Price Increase Treatment on the Price Sensitivity of the Decision to Punish by Faith in Intuition



Notes: Only high Faith in Intuition scores have a greater price sensitivity of punishment to a price increase than to a price decrease. All interaction terms between Faith in Intuition, punishment price and treatment dummy are added to the logit regression in column 1 of Table 2.4.

2.5 Discussion

2.5.1 Intuition and Pro-sociality

We find a positive but statistically insignificant relationship between individual reliance on intuition and giving. Previous studies have found conflicting results on whether prosociality is intuitive or not (Rand *et al.* , 2012; Tinghög *et al.* , 2013; Verkoeijen & Bouwmeester, 2014; Myrseth & Wollbrant, 2015). Our study contributes to this issue by showing a non-linear relationship between intuition and prosocial giving. As suggested by Figure 2.3, the most intuitive group still seem to be the most willing to give, but the frequency of giving is

roughly the same for the other three less intuitive groups. Subjects with little reliance on their own intuition might still give fairly for other reasons, such as a tendency to 'rely on the opinion and received wisdom of others' (Hodgkinson & Clarke, 2007). If we can control for the tendency to rely on others' opinion, we might find a clearer relationship between individual reliance on intuition and prosocial giving.

2.5.2 Deliberation and the Impulse to Punish

We find that subjects who tend to think more deliberately are less likely to punish than subjects who are less deliberate. The motivation for punishment might be spite or fairness concerns, depending on whether the case is when the first player gives zero, one or two dollars. Our finding suggests that deliberation restricts the impulse to punish in all three cases. A previous study also shows that the punishment rate drastically falls if decisions are delayed by around 10 minutes (Grimm & Mengel, 2011).

Carpenter & Matthews (2009) find that people often use different sets of norms in deciding whether to punish and how much to punish. Table 2.5 uses the level of punishment as the dependent variable, and the sign and significance of the estimates are similar to those in Table 2.4 except that the effect of deliberate thinking on punishment level is negative but not statistically significant. Further research is needed to investigate whether there is a difference between the effects of deliberation on the frequency and severity of punishment.

2.5.3 Intuition and Price Sensitivity of Punishment

We find that individual reliance on intuition is primarily associated with how sensitive the punishment decision is to changes in the price of punishment. In particular, more intuitive subjects are less price sensitive in terms of both the frequency and severity of punishment, suggesting a tendency to choose the same level of punishment at different prices of punishing.[4]

A possible explanation comes from the theory that moral judgment is the outcome of quick, automatic intuitions (Haidt, 2001; Haidt & Joseph, 2004;

[4]Whether the dependent variable is the decision to punish or the level of punishment, the interaction term between the price of punishment and Faith in Intuition is positive and significant.

Table 2.5: Determinants of the Second Player's Level of Punishment

Dependent Variable: Punishment ($)	OLS Model		
	If partner	If partner	If partner
Independent Variables	gives $0	gives $1	gives $2
High punishment price	-0.760***	-0.389***	-0.069*
	(0.078)	(0.054)	(0.039)
ln(*Need for Cognition*)	-0.593	-0.813***	-0.453**
	(0.407)	(0.285)	(0.220)
ln(*Faith in Intuition*)	-0.366	0.254	0.203
	(0.307)	(0.189)	(0.128)
Price Increase Treatment	0.054	-0.033	-0.033
	(0.152)	(0.111)	(0.088)
Female	-0.034	-0.083	-0.121
	(0.160)	(0.116)	(0.093)
White	-0.039	-0.184	-0.307***
	(0.182)	(0.124)	(0.101)
Bachelor's degree	0.076	0.034	0.045
	(0.164)	(0.115)	(0.087)
Cohabiting	-0.002	0.006	0.074
	(0.173)	(0.118)	(0.090)
Unemployed	0.170	-0.114	-0.091
	(0.213)	(0.144)	(0.106)
Household income < $25,000	0.247	0.124	0.089
	(0.173)	(0.121)	(0.100)
Age from 29 to 34	-0.006	-0.085	-0.210*
	(0.199)	(0.133)	(0.111)
Age from 35 to 49	-0.432**	-0.208	-0.293***
	(0.213)	(0.150)	(0.105)
Age at least 50	-0.030	0.044	-0.303**
	(0.290)	(0.203)	(0.128)
Constant	3.143***	2.114***	1.145***
	(0.727)	(0.503)	(0.406)
Subjects	325	325	325
Observations	650	650	650
R^2	0.091	0.079	0.097

Notes: The dependent variable is the level of punishment ranging from $0 to $4. The dummy variable *high punishment price* equals one when the price of punishing is 50¢ (versus low price of 10¢). The variable ln(*Need for Cognition*) is the natural logarithm of the degree to which subjects enjoy effortful analytic thinking on a scale from one to five. The variable ln(*Faith in Intuition*) is the logarithm of the degree to which subjects rely on intuitive feeling on a scale from one to five. The dummy variable *Price Increase Treatment* takes the Price Decrease treatment as the default. Robust standard errors in parentheses are clustered at the subject level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Graham *et al.* , 2012). Several studies have further suggested that intuitive thinking is associated with a preference for fairness and justice.[5] One theory of justice proposes that the offender deserves to be punished in proportion to the severity of the offense committed (e.g. Carlsmith *et al.* , 2002). Under this moral intuition, an analysis of the cost and benefit, such as the price of punishing, should not affect the decision to punish or the decision of how much to punish. As Kant (1790) argued, 'For justice would cease to be justice, if it were bartered away for any consideration whatever.' People who hold this moral intuition will not increase punishment simply because the price of punishment is lower. Therefore, the moral intuition of justice might be an explanation of why intuitive people tend to choose the same level of punishment, irrespective of cost.

In addition, we discover that the price sensitivity of punishing decision is greater in response to a price increase than to a price decrease for subjects with high reliance on intuition. In contrast, for subjects with low reliance on intuition, the decision to punish is similarly responsive to a price decrease and a price increase. The reason that highly intuitive subjects display differential sensitivity to price changes might be due to loss aversion relative to the first price of punishment as a reference point (Heidhues & Kőszegi, 2008). For instance, if the reference point is the low price of punishment in the first round, then the high price in the second round can generate a sense of loss, which amplifies the sensitivity to the price increase. Hence, people with loss aversion would punish in a way that responds to a price increase more than a price decrease, just as the highly intuitive subjects in our sample. However, to examine the impact of intuition on loss aversion, we should design a new experiment to test this directly.

2.6 Conclusion

Many experiments have shown that individuals differ considerably in their tendency to give or punish, emphasizing the importance to account for this heterogeneity (e.g. Andreoni & Miller, 2002; Anderson & Putterman, 2006;

[5] If moral intuition is innate, it must be found in brain responses and early in life. Children as young as three years old display signs that they dislike inequality when they received fewer gifts than other children did (LoBue *et al.* , 2011). A neuroimaging study showed that accepting unfair offers activated brain regions involved in self-control (Tabibnia *et al.* , 2008).

Carpenter, 2007). This study experimentally explores how individual differences in dual ways of processing information can explain this heterogeneity. Andreoni & Miller (2002) specifically argued that fairness and altruism must be addressed on an individual level because 'a model that predicts well in the aggregate may not help us understand the behaviour of individual actors.' Our study improves our understanding by showing that deliberate thinking decreases the tendency to punish and that reliance on intuition has a non-linear relationship with giving. Moreover, high reliance on intuition is associated with being less sensitive to punishment price changes and the asymmetry of responding to a price increase versus decrease. These key findings demonstrate that the measures of Faith in Intuition and Need for Cognition are a useful tool in understanding fairness and altruism that motivate giving and punishment.

2.7 Appendix A: Experimental Instructions

PARTICIPATION AGREEMENT

In order to participate in this research study, it is necessary that you give your informed consent. By responding you are indicating that you understand the nature of the research study and your role and that you agree to participate in the research. Please consider the following points before continuing:

- I understand that I am participating in academic research conducted by the University of Warwick.
- I understand the research team will use anonymized data in any presentations of research results. Data will not be associated with individuals except where necessary to pay rewards based on performance, and any identifying data will then be destroyed.
- I understand that my participation in this study is voluntary, and that after the study data collection has begun, I may refuse to participate further without any penalty.
- By continuing I am stating that I am at least 18 years of age and that I have read the above information and consent to participate in this study being conducted.

Please check this box to agree that you have read and understood the information above:

☐ I Agree

Payments

This survey will take about 15 minutes. Upon satisfactorily completing the survey, your work will be approved within 48 hours and you will receive $1.50.

You will be randomly matched with two other workers. Depending on your decisions and their decisions, you can earn an additional payment up to $20. It will be shown to you exactly how your additional payment is computed.

Only 1 worker who has satisfactorily completed the survey will be randomly selected to receive his/her additional payments. Any additional payment will be paid when we finish the analysis.

Instructions

You will be randomly matched with another worker, called Worker A. He/she will decide how to split $4 between you and himself/herself. You will then have an option to reduce his/her earnings.

If you choose to reduce Worker A's earnings, it will cost you something. The cost of reducing $1 of Worker A's earnings is 50 cents to you. For example, if you choose to reduce $2.50 of Worker A's earnings, the total cost of reduction to you is 0.5*2.50=1.25.

The experimenter will give you $2 before you decide how much you will reduce Worker A's earnings. You can either use them to cover the cost of reduction or keep them as part of your payment.

Worker A has three options of how to split $4: giving you $0, $1, or $2. You will need to indicate your choice of reduction for each of the three options. Since only one will be selected by Worker A, only in this specific case both of your choices will affect payments. Your earnings from interacting with Worker A are computed as the following:

Your earnings = (Amount you receive from Worker A) + $2 - (0.5)(How much you choose to reduce Worker A's earnings).*

Please allow us to check if we have explained things clearly. Below are some possible scenarios of your interaction with Worker A. Please tell us what your total cost of reduction is and what your earnings are for each scenario.

	What is your total cost of reduction?			What are your earnings?		
	$0.5	$1.25	$2	$0	$1.75	$3.5
If Worker A gives you $2 and you reduce $1 of A's earnings,	○	○	○	○	○	○
If Worker A gives you $1 and you reduce $2.5 of A's earnings,	○	○	○	○	○	○
If Worker A gives you $0 and you reduce $4 of A's earnings,	○	○	○	○	○	○

Recall: Your earnings are computed as the following:

Your earnings = (Amount you receive from Worker A) + $2 - (0.5)(How much you choose to reduce Worker A's earnings).*

Please indicate your choices in three cases below.

	How much would you like to reduce Worker A's earnings?	How angry are you with Worker A? Please rate on a scale of 1-7.
In the case where Worker A gives you $0 out of $4,	------	------
In the case where Worker A gives you $1 out of $4,	------	------
In the case where Worker A gives you $2 out of $4,	------	------

You will interact with Worker A again with the same procedure as before. The only difference is that the cost of reducing $1 of Worker A's earnings is now 10 cents. Therefore, your earnings are computed as the following:

Your earnings = (Amount you receive from Worker A) + $2 - (0.1)(How much you choose to reduce Worker A's earnings).*

Please indicate your choices in the three cases below.

	How much would you like to reduce Worker A's earnings?	How angry are you with Worker A? Please rate on a scale of 1-7.
In the case where Worker A gives you $0 out of $4,	------	------
In the case where Worker A gives you $1 out of $4,	------	------
In the case where Worker A gives you $2 out of $4,	------	------

Now you will be randomly matched with a different worker, called Worker B. You will take the opposite role. You can decide how to split $4 between you and him/her. Worker B will then have an option to reduce your earnings. Worker B's cost of reducing $1 of your earnings is 50 cents. Your earnings from interacting with Worker B are computed as the following:

Your earnings = $4 - (Amount you give to Worker B) - (How much Worker B reduces your earnings).

How much would you like to give to Worker B?

- $0 out of $4.
- $1 out of $4.
- $2 out of $4.

You will interact with Worker B again with the same procedure as before. The only difference is that Worker B's cost of reducing $1 of your earnings is 10 cents. Your earnings are computed as the following:

Your earnings = $4 - (Amount you give to Worker B) - (How much Worker B reduces your earnings).

How much would you like to give to Worker B?

- $0 out of $4.
- $1 out of $4.
- $2 out of $4.

Your additional payment is your total earnings from interacting with Worker A and your total earnings from interacting with Worker B. If you are selected, you will receive your additional payment based on the Worker A and Worker B's actual decisions.

2.8 Appendix B: Questionnaire

We used the Rational-Experiential Inventory questionnaire developed by Pacini and Epstein (1999). It has 20 items for the Need for Cognition scale and 20 items for the Faith in Intuition scale. Participants were asked to rate how strongly they agree with each item on a five-point scale ranging from 1

(*strongly disagree*) to 5 (*strongly agree*). To create a score for Need for Cognition or Faith in Intuition, the twenty responses for that scale were added up and divided by twenty.

Need for Cognition (20 items)

1. I have a logical mind.
2. I prefer complex problems to simple problems.
3. I am not a very analytical thinker. (Recode)
4. I try to avoid situations that require thinking in depth about something. (Recode)
5. I don't reason well under pressure. (Recode)
6. Thinking hard and for a long time about something gives me little satisfaction. (Recode)
7. I am much better at figuring things out logically than most people.
8. I usually have clear, explainable reasons for my decisions.
9. Thinking is not my idea of an enjoyable activity. (Recode)
10. I have no problem thinking things through carefully.
11. Learning new ways to think would be very appealing to me.
12. I'm not that good at figuring out complicated problems. (Recode)
13. I enjoy intellectual challenges.
14. Reasoning things out carefully is not one of my strong points. (Recode)
15. I enjoy thinking in abstract terms.
16. Using logic usually works well for me in figuring out problems in my life.
17. I don't like to have to do a lot of thinking. (Recode)
18. Knowing the answer without having to understand the reasoning behind it is good enough for me. (Recode)
19. I am not very good at solving problems that require careful logical analysis. (Recode)
20. I enjoy solving problems that require hard thinking.

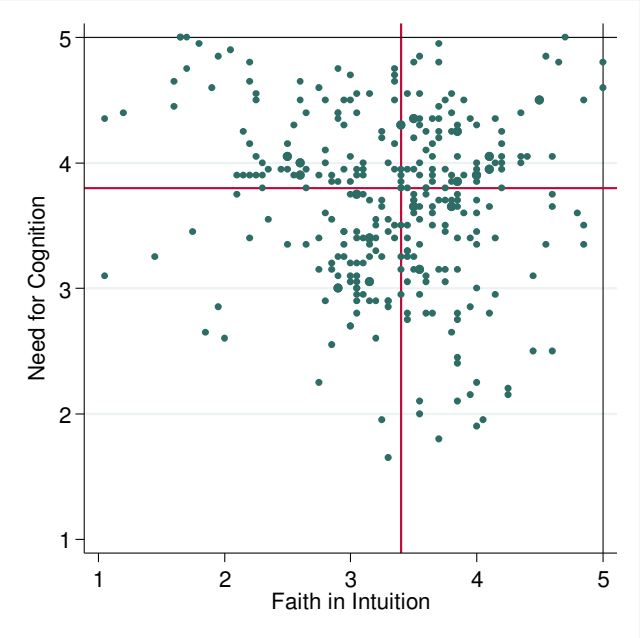
Faith in Intuition (20 items)

1. I believe in trusting my hunches.
2. I trust my initial feelings about people.
3. I like to rely on my intuitive impressions.
4. I don't like situations in which I have to rely on intuition. (Recode)

5. Intuition can be a very useful way to solve problems.
6. I would not want to depend on anyone who described himself or herself as intuitive. (Recode)
7. I don't think it is a good idea to rely on one's intuition for important decisions. (Recode)
8. When it comes to trusting people, I can usually rely on my gut feelings.
9. I can usually feel when a person is right or wrong, even if I can't explain how I know.
10. I hardly ever go wrong when I listen to my deepest gut feelings to find an answer.
11. I think it is foolish to make important decisions based on feelings. (Recode)
12. I tend to use my heart as a guide for my actions.
13. I often go by my instincts when deciding on a course of action.
14. I generally don't depend on my feelings to help me make decisions. (Recode)
15. I think there are times when one should rely on one's intuition.
16. Using my gut feelings usually work well for me in figuring out problems in my life.
17. I don't have a very good sense of intuition. (Recode)
18. If I were to rely on my gut feelings, I would often make mistakes. (Recode)
19. I suspect my hunches are inaccurate as often as they are accurate. (Recode)
20. My snap judgements are probably not as good as most people's. (Recode)

2.9 Appendix C: Additional Analysis

Figure 2.9: No correlation between Need for Cognition and Faith in Intuition



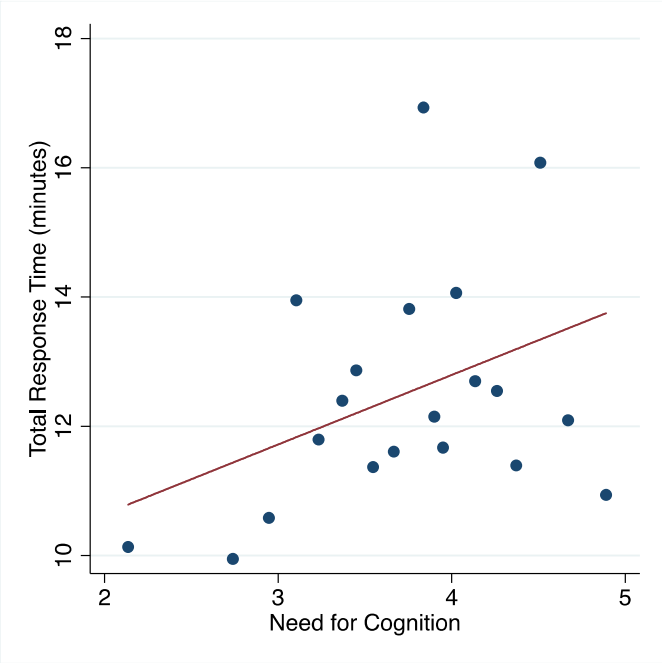
Notes: The size of each dot represents the number of subjects. The median score for Faith in Intuition is 3.4, and the median score for Need for Cognition is 3.8.

Table 2.6: Distribution of Punishing Behaviour in the Case of Receiving zero dollars (%)

		High Cost of Punishment (50¢)	
		Not Punish	Punish
Low Cost of Punishment (10¢)	Not Punish	28.6	0.3
	Punish	15.7	55.4

Notes: In response to player 1 giving zero dollars, each subject makes two punishment decisions, and the cost of punishment is high in one round and low in another. Each cell indicates the percentage of subjects who make a particular pair of decisions to punish. For example, 28.6% of subjects choose not to punish under both conditions.

Figure 2.10: Deliberate People Are Slower Decision Makers



Notes: This binned scatter plot shows that relationship between total response time and Need for Cognition. Subjects are grouped into twenty equal-sized bins based on their Need for Cognition scores, and each bin indicates its average Need for Cognition and its average response time, measured in minutes.

Chapter 3

A Theory of the Efficiency of Divorce with Interdependent Preferences

3.1 Introduction

When does the possibility to divorce benefit or harm a society? According to Becker (1991), 'A husband and wife would both consent to a divorce if, and only if, they both expected to be better off divorced.' The key assumption is that spouses can bargain efficiently within marriage.[1] However, American household data revealed that spouses have private information that can create inefficient bargaining. Friedberg & Stern (2014) used unique questions from the National Survey of Families and Households to show that more than half of spouses have misperceptions about their partners' happiness after a hypothetical divorce. In particular, their data indicated that the divorce rate was higher for couples where one spouse overestimated how unhappy the partner would be if they separated than for couples where one spouse had the correct perception about their partner. Moreover, Zhylyevskyy (2012) used spouses' reported opinions to infer their true value of the outside option probabilistically and estimated a structural model to predict the frequency of inefficient divorces. The evidence showed that 22.6% of divorces are inefficient because these divorced couples would have been better off if they stayed married.

[1] This is an extension of the Coase theorem (1960) and is also adopted by McElroy & Horney (1981), Chiappori (1988), Browning & Chiappori (1998), Basu (2006), and Mazzocco (2007).

This chapter develops a model to explore the welfare consequences of allowing for divorce (versus no possibility of divorce) in the presence of asymmetric information. Suppose each spouse has a private value of the marriage and can signal one's value through the process of bargaining. When divorce is allowed, spouses can negotiate non-cooperatively within marriage by using the threat of divorce.[2] A simple bargaining procedure is used to capture the effects of two-sided asymmetric information. The setting begins with a married couple, in which one player is given the opportunity to initiate bargaining or not. If the first player chose not to initiate bargaining, the marriage continues as it is. If bargaining were initiated, the second player can propose either to give a transfer to the first spouse or demand a payment from him or her. Then the first player can choose either to accept the proposed gift/request and stay in marriage or to reject and end in divorce.

This model can be broadly applied to any kind of partnership, such as business or employment relations, where there is asymmetric information about the other partner's value of partnership. Dissolution can be beneficial, but often it is jointly suboptimal because partners might bargain inefficiently with the threat of separation. Thus, sometimes a fully committed partnership can be better than partnerships that have the opportunity to freely bargain and dissolve. All the results of this chapter can be applied to other types of partnership, but in the following analysis we use marital relationship as an example.

The first result is that the expected welfare of getting married with a possibility to divorce is ex ante worse than with no possibility to divorce, if spouses hold a sufficiently strong prior belief that their partner has a high value of the marriage. The reason is that both spouses try to extract too much surplus from each other, leading some efficient marriages, that are jointly optimal, to divorce. If spouses do not hold such prior belief, the expected welfare of getting married with a possibility to divorce is ex ante better because of its ability to end inefficient marriages, that are jointly sub-optimal.

We model interdependent preferences by assuming that spouses may care about their evaluation of their partner's hidden value of the marriage. This means that one's high or low value of the marriage will affect the other's

[2]Spouses became less cooperative, when the divorce cost fell due to the move from mutual consent to unilateral divorce laws across US states. Evidence shows that both the investment in marriage-specific capital (Stevenson, 2007) and the degree of risk-sharing drop (Halla & Scharler, 2012).

marital utility. Evidence suggests that married couples have interdependent preferences and thus 'justifies incorporating "love" into economic theory' (Friedberg & Stern, 2014). Our aim is to investigate how the degree to which spouses care about their partner's value of the marriage will impact the welfare analysis of allowing for divorce. Since one's action during the bargaining process can reveal one's high or low value of the marriage, the action can also increase or decrease the partner's overall marital utility, possibly generating an outcome of restoring an initially inefficient marriage.

This type of interdependent utility influences marital bargaining in several interesting ways. As players care more about their partner's value of the marriage, a second player who has a high value will give a smaller transfer, in equilibrium, to a first player who initiated bargaining. The reason is that this transfer can reveal the second player's high value of the marriage and thus raise the partner's marital utility (and the willingness to accept a smaller transfer). This can also indirectly reduce the chances of inefficient bargaining and of ending a jointly optimal marriage. Consequently, the ex ante expected welfare of getting married with a possibility to divorce increases as players care more about their partner's value of the marriage.

Moreover, as players have a sufficiently high concern for their partner's value of the marriage, a second player who has a high value will even demand a payment from a first player who has a low value. It may seem surprising that in equilibrium the first player would still accept the demand. The reason is that this demanded payment is not as harsh as what a second player who has a low value would demand and thus reveals the high value of the marriage, which raises the first player's utility and confidence in the marriage. This suggests that the signal or information conveyed through an action can have a positive impact, even though the action does not seem friendly.

The main contribution is that incorporating interdependent preferences can eliminate the possibility of inefficient divorce generated by asymmetric information. An increase in the degree of caring about the partner's private value of marriage generally improves and never reduces the ex ante welfare of getting married with a possibility to divorce. The increased welfare is gained by ending inefficient marriages and sometimes restoring an initially inefficient marriage. When players care sufficiently about their partner's value of the marriage, the possibility to divorce yields an ex ante welfare that is at least as high as with no possibility to divorce. In contrast, only when players care

little about their partner's value and hold a sufficiently high prior belief that their partner has a high value, allowing for divorce can yield lower ex ante welfare of getting married than if there is no possibility to divorce.

This research contributes to a large literature that has studied the impact of changes in divorce laws on various outcomes, including divorce rate (Wolfers, 2006), domestic violence (Stevenson & Wolfers, 2006), marriage-specific capital (Stevenson, 2007), trends in divorce rate (Matouschek & Rasul, 2008), and non-contractible marital investments (Wickelgren, 2009). Our study focuses on two versions of divorce law, which either allows for divorce or not. This choice of design makes the model more tractable in analysing the impact of bargaining with asymmetric information on the expected welfare of getting married.

The remainder of the chapter is organized into four sections. Section 3.2 lays out the model and key definitions. Section 3.3 analyses the case in which spouses do not care about their partner's value of the marriage and the welfare-implications of allowing for divorce. Section 3.4 analyses the case in which spouses can care about their partner's value of the marriage up to the same degree to which they care about their own value of the marriage. Section 3.5 concludes. All proofs are in the Appendix.

3.2 Model

Consider a married couple who observe each other's material gains from marriage, such as labour specialisation, economies of scale and risk sharing. However, each spouse has a hidden valuation of the marriage. Suppose each spouse $i \in \{1, 2\}$ receives an equal material gain from being married $z > 0$. There are two types of spouses, denoted by $\mu_i \in \{H, -L\}$ where $H, L > 0$. A spouse who has a high valuation of the marriage, $\mu_i = H > 0$, is a high type, while a spouse who has a low valuation of the marriage, $\mu_i = -L < 0$, is a low type. Spouses have a common prior belief that the other spouse is a high type with probability $q \in [0, 1]$.

The model incorporates interdependent preferences by assuming that each spouse cares not only about one's own value of the marriage but also the partner's value of the marriage. This means that one is more satisfied with the marriage if one knows or believes that the partner has a high valuation of the marriage. To formalise this, let $\hat{\mu}_{ij} = E_i[\mu_j|h]$ denote player i's expected

evaluation of player j's value of the marriage according to the history of the observed actions, which is denoted by h.[3] Now the utility function of being married is

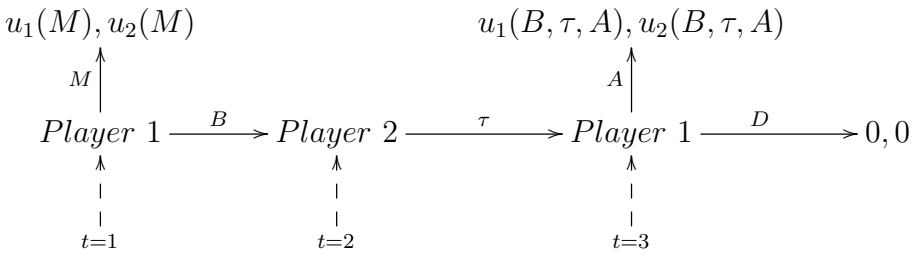
$$u_i = z + \mu_i + \theta\hat{\mu}_{ij},$$

where $\theta \in [0, 1]$ is the degree to which one cares about the partner's value of the marriage. When $\theta = 0$, both players do not care about each other's private value. When $\theta = 1$, both spouses care about each other's private value to the same degree as they care about their own value. For simplicity, let the divorce payoff be zero for both spouses.[4] Whether a marriage is efficient or not depends on the joint utility of being married and the joint utility of divorce.

Definition 1. *A marriage is efficient whenever* $u_1 + u_2 \geq 0$ *and is inefficient whenever* $u_1 + u_2 < 0$.

Note that this definition depends on the sum of the two players' outcomes, not the distribution of outcomes. By implication, a divorce is inefficient if and only if it ends an efficient marriage.

Figure 3.1: Bargaining Procedure



When a married couple has a possibility to negotiate with the divorce threat, they proceed to a three-stage sequential bargaining game (see Figure 3.1). In the first stage, player 1 can choose to initiate bargaining (B) or to

[3]This modelling choice is related to Ellingsen & Johannesson (2008), who developed a model in which the agent's utility of esteem depends on the pro-sociality of the principal. An agent cares more about social esteem from a good principal than from a bad principal. Similarly, the spouse in our model cares more about having a partner who has a high value of the marriage than one who has a low value of the marriage.

[4]If the divorce payoffs favour one spouse or the other, the bargaining power derived from the outside options can change the equilibrium behaviour. For example, Rainer (2007) explains how the legal divorce reform from equitable distribution (ruled by courts according to the contributions made to the marital assets by the spouses and their future needs) to equal division may have increased the demand of prenuptial contracts.

stay married (M). Denote the set of pure actions of player 1 of type μ_1 by $\mathcal{A}_{\langle 1,\mu_1\rangle} = \{M, B\}$ and a generic action by $a_{\langle 1,\mu_1\rangle}$. If player 1 refrains from initiating bargaining and chooses M, the marriage continues and the utility payoffs to players 1 and 2 are, respectively,

$$u_{\langle 1,\mu_1\rangle}(M) = z + \mu_1 + \theta\hat{\mu}_{12} \quad \text{and} \quad u_{\langle 2,\mu_2\rangle}(M) = z + \mu_2 + \theta\hat{\mu}_{21}.$$

If player 1 chooses B, the game proceeds to the second stage in which player 2 of type μ_2 can propose a transfer of utility, $a_{\langle 2,\mu_2\rangle} = \tau \in \mathbb{R} = \mathcal{A}_{\langle 2,\mu_2\rangle}$, which can be positive or negative. If the transfer is positive, player 2 proposes to give a payment to player 1. If the transfer is negative, player 2 places a demand on player 1. In the third stage, player 1 can choose either to accept the gift/request and continue the marriage (A) or to reject the proposal and end in a divorce (D). The utility payoffs to players 1 and 2 when the transfer is accepted are, respectively,

$$\begin{aligned} u_{\langle 1,\mu_1\rangle}(B,\tau,A) &= z + \mu_1 + \theta\hat{\mu}_{12} + \tau \text{ and} \\ u_{\langle 2,\mu_2\rangle}(B,\tau,A) &= z + \mu_2 + \theta\hat{\mu}_{21} - \tau. \end{aligned}$$

3.3 The Case Without Interdependent Utility

We first examine the case in which spouses do not care about each other's private value of the marriage, that is, $\theta = 0$. Since each player can be of high or low type, there are four possible combinations: HH, HL, LH, LL, where the first letter refers to the type of the first player and the second letter refers to the type of the second player. If $z > L$, all four combinations are efficient marriages and there will be no need of divorce. Thus, we make the following assumption to ensure that some marriages are efficient while others are not.

Assumption 1. $z < L < 2z + H$.

This assumption implies that player i with a low valuation of the marriage will prefer divorce because $u_{\langle i,L\rangle} = z - L < 0$. Thus, a marriage between two low types (LL) is inefficient. However, a marriage between a low type and a high type (HL, LH) is jointly optimal because $u_{\langle i,H\rangle} + u_{\langle j,L\rangle} = 2z + H - L > 0$. A marriage of two high types (HH) is for sure better off staying together.

3.3.1 Equilibrium Analysis

When the possibility to negotiate and divorce is present, there exists two classes of equilibria depending on the value of q, the prior belief that the partner has a high valuation of the marriage. The following presents the class of separating equilibria:

Proposition 1. *If $\theta = 0$ and $0 \leq q \leq \hat{q}$, there exists a separating equilibrium, where $\hat{q} \equiv \frac{z+H}{H+L}$:*

> *At $t = 1$, a high-type player 1 chooses to stay married (M) and a low-type player 1 chooses to initiate bargaining (B).*
>
> *At $t = 2$, a high-type player 2 proposes to give a transfer $\tau_H = -z + L$ and a low-type player 2 proposes to demand a transfer $\tau_L = -z - H$. Player 2's posterior belief: $\Pr(\mu_1 = H|M) = 1$ or $\Pr(\mu_1 = H|B) = 0$.*
>
> *At $t = 3$, a high-type player 1 accepts both τ_H and τ_L, and a low-type player 1 accepts τ_H but rejects τ_L. Player 1's posterior belief: $\Pr(\mu_2 = H|\tau_H) = 1$ or $\Pr(\mu_2 = H|\tau_L) = 0$.*

Proof. See the Appendix. □

In the separating equilibrium, the first players of different types can fully reveal their private value of the marriage through their decision at period one. Player 1 who has a low value of the marriage will always initiate bargaining using the threat of divorce (B) because they prefer divorce outcome to their marital payoffs. Player 1 who has a high value of the marriage will always refrain from initiating bargaining (M) because their marital utility is higher than the expected utility of initiating bargaining.

If bargaining is initiated at period one, the second players also reveal their value of the marriage by choosing different strategies in proposing a transfer at period two. Player 2 who has a high valuation of the marriage would propose to give player 1 a positive payment $\tau_H = -z + L$, which is the minimum that a low-type player 1 would accept and continue the marriage. Player 2 who has a low valuation of the marriage would not give anything but demand player 1 to give a payment $\tau \leq z - L$ so that remaining in marriage would be at least as good as the divorce outcome. Any demanded transfer will for sure be rejected by a low-type player 1, but a high-type player 1 will accept a proposal if and only if $\tau \geq -z - H$. Thus, the optimal strategy for

a low-type player 2 under any possibilities of meeting a high-type player 1 is to demand $\tau_L = -z - H$. At period three, a high-type player 1 is willing to accept either a gift τ_H or a demand τ_L, but a low-type player 1 is only willing to accept a gift τ_H.

In this class of separating equilibria, the choice to initiate bargaining by player 1 always leads to jointly optimal outcomes by ending inefficient marriages and preserving efficient marriages. For instance, a marriage between two low-type players is jointly sub-optimal, and in equilibrium this couple does end up in divorce. A marriage between a low-type player 1 and a high-type player 2 is jointly optimal, and in equilibrium this couple does remain in marriage after player 2 gives a positive transfer to compensate player 1.

The condition for the existence of separating equilibrium is that a high-type player 1 must prefer staying married (M) to initiating bargaining (B) in expectation: $E\left(u_{\langle 1,H\rangle}(M)\right) \geq E\left(u_{\langle 1,H\rangle}(B)\right) \Rightarrow \quad z + H \geq q(z + H + \tau_H)$

$$\Rightarrow \quad q \leq \frac{z+H}{H+L} = \hat{q}. \tag{3.1}$$

If q, the prior belief that the partner has a high valuation of the marriage, is above the threshold $\hat{q}$, a high-type player 1 would play a mixed strategy at period one. To initiate bargaining can yield a higher payoff if he meets a high-type player 2 who is willing to give a positive transfer. However, there is also a risk in initiating bargaining because the high-type player 1 will have to accept a demand placed on him if he meets a low-type player 2. The chances of facing this risk are low if q is sufficiently high, that is, if the prior belief suggests that player 2 is not likely to be a low type. The following result characterises the mixed strategy equilibrium:

Proposition 2. *If $\theta = 0$ and $\hat{q} < q < 1$, there exists a mixed strategy equilibrium, where $\hat{q} = \frac{z+H}{H+L}$:*

> *At $t = 1$, a high-type player 1 chooses to stay married (M) with probability p and chooses to initiate bargaining (B) with probability $1-p$, and a low-type player 1 always initiates bargaining (B).*
>
> *At $t = 2$, a high-type player 2 proposes to give $\tau_H = -z + L$ with probability r and to demand $\tau_L = -z - H$ with probability $1 - r$, and a low-type player 2 always demands $\tau_L = -z - H$. Player 2's posterior belief: $\Pr(\mu_1 = H|M) = 1$ or $\Pr(\mu_1 = H|B) = \frac{q(1-p)}{1-qp}$.*

At $t = 3$, a high-type player 1 accepts both τ_H and τ_L, and a low-type player 1 accepts τ_H but rejects τ_L. Player 1's posterior belief: $\Pr(\mu_2 = H|\tau_H) = 1$ *or* $\Pr(\mu_2 = H|\tau_L) = \frac{q(1-r)}{1-qr}$.

The mixing probabilities are $p \equiv \frac{q(2z+2H)-(2z+H-L)}{q(H+L)}$ *and* $r \equiv \frac{z+H}{q(H+L)}$.

In response to a high-type player 1's mixed strategy, a player 2 who has a high valuation of the marriage also chooses a mixed strategy in equilibrium. The high-type player 2 mixes between giving a positive transfer $\tau_H = -z + L$ (which can persuade a low type to accept and remain in the marriage) and demanding a negative payment $\tau_L = -z - H$ (which can extract the maximum surplus from a high type). Therefore, in equilibrium, a high type player 1 chooses to stay married with probability p and to initiate bargaining with probability $1-p$ at period one in order to make a high-type player 2 indifferent between giving τ_H and demanding τ_L at period two. In return, the high-type player 2 chooses to give τ_H with probability r and to demand τ_L with probability $1 - r$ in order to make the high-type player 1 indifferent between initiating and not initiating bargaining.

These mixed strategies can cause a divorce to be a sub-optimal outcome. If a low-type player 1 meets a high-type player 2 and this player 2 'happens' (with probability $1 - r$) to demand a negative transfer τ_L, the low-type player 1 will reject the proposal and end in divorce. However, this type of marriage is better off staying together because their joint utility is $u_1+u_2 = 2z+H-L > 0$, indicating an efficient marriage. Hence, the divorce outcome in this case is inefficient.

Claim 1. *If $\theta = 0$, there is no pooling equilibrium.*

Here we provide the intuition of this claim. Since a low-type player 1 will always initiate bargaining, the only possible pooling equilibrium is when both types of player 1 choose to initiate bargaining at period one. Suppose there exists such a pooling equilibrium. Then, a player 1 cannot signal his or her type at period one, so player 2's posterior belief about player 1's type is the same as the prior belief q. Now a low-type player 2 always demands τ_L because it's the weakly dominant strategy. If a high-type player 2 also demands τ_L, a high-type player 1 will then deviate from initiating bargaining because bargaining only leads to a lower payoff. Thus, this set of strategies cannot constitute an equilibrium. If a high-type player 2 chooses to give τ_H,

the chances of meeting a low type must be sufficiently high so that giving τ_H would the optimal strategy in preserving marriage with a low type. However, for player 1, because the risk of meeting a low type who demands a payment is now higher than the potential gain of meeting a high type who gives a payment, a high-type player 1 will again deviate from initiating bargaining. Therefore, there is no pooling equilibrium in which both types of player 1 choose to initiate bargaining at period one.

3.3.2 Welfare Analysis

Now we investigate the conditions in which having the possibility to bargain with the threat of divorce yields a higher or lower ex ante expected welfare than having no possibility for divorce. Suppose ex ante the probability of having a high valuation of the marriage is q. In the bargaining procedure, it is equally likely to be the first or second player. Let W_N denote a player's ex ante expected welfare of getting married when there is no possibility for negotiation or divorce. Then it would simply be the expected marital utility:

$$W_N = q(z + H) + (1 - q)(z - L).$$

When there is the possibility to negotiate with divorce threat, the bargaining procedure results in a separating equilibrium if the prior belief that any player is a high type is sufficiently low. Let W_S denote a player's ex ante expected welfare of getting married in the separating equilibrium. In this equilibrium, the possibility to initiate bargaining leads an inefficient marriage between two low-type players to divorce and keeps all other efficient marriages together. Compared to W_N, the possibility of divorce yields an expected welfare gain by eliminating the negative expected payoff of being in a marriage between two low-type players $(1 - q)^2(z - L)$. Hence,

$$W_S = q(z + H) + q(1 - q)(z - L) \quad \text{if } 0 \leq q \leq \hat{q}. \tag{3.2}$$

We have also seen that the bargaining procedure results in a mixed strategy equilibrium if q is sufficiently large. One potential outcome is to divorce an efficient marriage in which a low-type player 1 meets a high-type player 2 who 'happens' (with probability $1 - r$) to demand a negative transfer from player 1 who in turn rejects it. Thus, the ex ante expected welfare if

$\hat{q} < q \leq 1$ is

$$W_X = q(z+H) + q(1-q)(z-L) \underbrace{- \left(\frac{1}{2}\right)[(1-q)q(1-r)(2z+H-L)]}_{\text{L meeting H who chooses } \tau_L}$$

The first two terms are the same as the expected welfare in the separating equilibrium. The third term is the expected loss caused by ending an efficient marriage described above. Next, we investigate whether the loss of divorcing jointly optimal marriages can ever outweigh the gain of efficient divorces.

Proposition 3. *The difference in the ex ante expected welfare of getting married between having and not having a possibility to bargain with divorce threat depends on a threshold* $\tilde{q} \equiv 1 - \frac{(L-z)(2z+H-L)}{(H+L)^2}$.

$W_S - W_N > 0$ *if* $0 \leq q \leq \hat{q}$;

$W_X - W_N > 0$ *if* $\hat{q} < q < \tilde{q}$;

$W_X - W_N = 0$ *if* $q = \tilde{q}$ *or* $q = 1$;

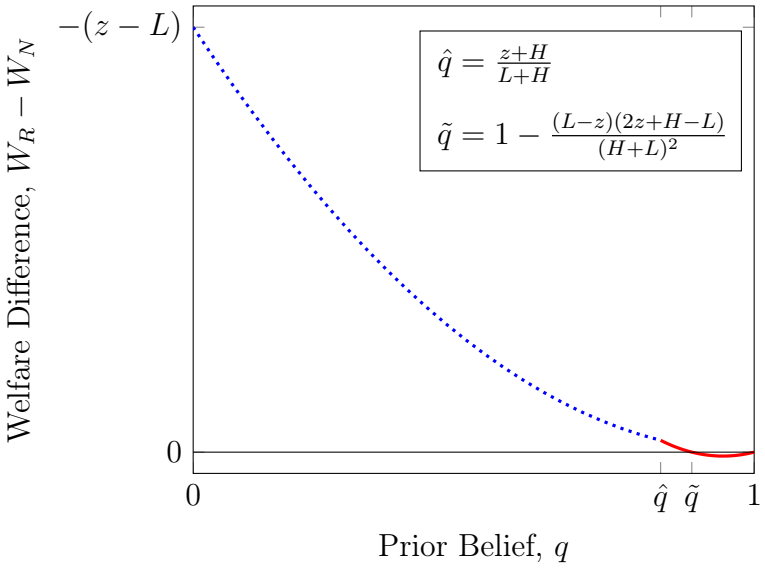
$W_X - W_N < 0$ *if* $\tilde{q} < q < 1$;

Figure 3.2 illustrates that the expected welfare difference between having and not having the opportunity of divorce decreases in q if $0 \leq q \leq \hat{q}$ (see the dotted line). The reason is that the gain from ending an inefficient marriage between two low-type players becomes smaller when players are ex ante more likely to be a high type. However, if $q > \hat{q}$, the expected welfare difference can fall below zero, as the solid line in Figure 3.2 indicates. This is due to two effects of having a higher q:

$$W_R - W_N = \underbrace{-(1-q)^2(z-L)}_{\text{Gain of ending LL}} \underbrace{- \left(\frac{1}{2}\right)[(1-q)q(1-r)(2z+H-L)]}_{\text{Loss of ending LH who chooses } \tau_L}.$$

The first term is the gain from ending an inefficient marriage through divorce, and the second term is the loss from divorcing a jointly optimal marriage through inefficient bargaining. As players are more likely to have a high value of the marriage, the chances of inefficient marriage fall. Moreover, the expectation that the partner has a high value of the marriage drives people to bargain inefficiently.

Figure 3.2: Welfare Analysis



3.4 The Case With Interdependent Utility

Now we consider the general case in which spouses can care about their evaluation of their partner's value of the marriage to varying degrees. Player i's marital utility is then $u_i = z + \mu_i + \theta E_i[\mu_j|h]$, where $0 \leq \theta \leq 1$ is the degree of i's concern for j's value of the marriage and h is the history of j's actions. For simplicity, we assume $H = L = k$, meaning that a high value of the marriage has the same magnitude as a low value of the marriage, and thus $\mu_i = \{k, -k\}$. Relative to z the material gains from marriage, k represents the magnitude of the effect of personal value of the marriage on marital utility. The relation between k and z is important in determining whether a marriage is jointly optimal or not. Similar to Assumption 1, we make the following assumption to ensure that at least a marriage between two low-type spouses is inefficient (if $\theta E_i[\mu_j|h] = 0$).

Assumption 2. $z < k < 2z$.

3.4.1 Equilibrium Analysis

We first discuss the case when $0 \leq \theta < \frac{k-z}{k}$ and then the case when $\frac{k-z}{k} \leq \theta \leq 1$. The first case is similar to the previous section in generating both separating and mixed strategy equilibria (Propositions 4 and 5). The second

case generates separating equilibria as well as a new class of pooling equilibria (Propositions 6 and 7).

The following presents the separating equilibrium if players care relatively little about their partner's value of the marriage and if they have a sufficiently low q, the prior belief that their partner is a high type.

Proposition 4. *If* $0 \leq \theta < \frac{k-z}{k}$ *and* $0 \leq q \leq \hat{q}(\theta)$*, there exists a separating equilibrium, where* $\hat{q}(\theta) \equiv \frac{z+k-\theta k}{2k-2\theta k}$*:*

> *At* $t = 1$*, a high-type player 1 stays married* (M) *and a low-type player 1 initiates bargaining* (B)*.*
>
> *At* $t = 2$*, a high-type player 2 gives* $\tau_H(\theta) = -z + k - \theta k$ *and a low-type player 2 demands* $\tau_L(\theta) = -z - k + \theta k$*. Player 2's posterior belief:* $\Pr(\mu_1 = H|M) = 1$ *or* $\Pr(\mu_1 = H|B) = 0$*.*
>
> *At* $t = 3$*, a high-type player 1 accepts both* $\tau_H(\theta)$ *and* $\tau_L(\theta)$*, and a low-type player 1 accepts* $\tau_H(\theta)$ *but rejects* $\tau_L(\theta)$*. Player 1's posterior belief:* $\Pr(\mu_2 = H|\tau_H) = 1$ *or* $\Pr(\mu_2 = H|\tau_L) = 0$*.*

In equilibrium, a high-type player 2 would give a positive payment $\tau_H(\theta) = -z + k - \theta k$ to compensate a low-type player 1, who is willing to accept it. A key observation is that $\tau_H(\theta)$ is decreasing in θ. This means that as players care more about their partner's value of the marriage, a low-type player 1 is willing to accept a smaller payment. The reason is that the low-type player 1 becomes more satisfied with the marriage not only because of the amount of $\tau_H(\theta)$ but also because of its ability to raise his belief that player 2 is a high type. Therefore, a high-type player 2 can give less as θ increases.

In contrast, a low-type player 2 would demand a negative transfer $\tau_L(\theta) = -z - k + \theta k$ from a high-type player 1, who is willing to meet the demand. Here the absolute value of $\tau_L(\theta)$ is decreasing in θ. This means that as players care more about their partner's value of the marriage, a high-type player 1 is only willing to meet a smaller demand. The reason is that the high-type player 1 becomes less satisfied with the marriage when $\tau_L(\theta)$ reveals that player 2 is a low type. Therefore, a low-type player 2 can demand less as θ increases.

The following proposition specifies the mixed strategy equilibrium that exists if q the prior belief that the partner is a high type becomes sufficiently high and players care relatively little about their partner's type.

Proposition 5. *If* $0 \leq \theta < \frac{k-z}{k}$ *and* $\hat{q}(\theta) < q < 1$*, there exists a mixed strategy equilibrium, where* $\hat{q}(\theta) = \frac{z+k-\theta k}{2k-2\theta k}$*:*

At $t = 1$*, a high-type player 1 stays married* (M) *with probability* $p(\theta)$ *and initiates bargaining with probability* $1 - p(\theta)$*, and a low-type player 1 initiates bargaining* (B)*.*

At $t = 2$*, a high-type player 2 gives* $\tau_H(\theta) = -z + k - \theta k$ *with probability* $r(\theta)$ *and demand* $\tau_L(\theta) = -z - k - \theta E_1[\mu_2|\tau_L]$ *with probability* $1 - r(\theta)$*, and a low-type player 2 demands* $\tau_L(\theta) = -z - k - \theta E_1[\mu_2|\tau_L]$*. Player 2's posterior belief:* $\Pr(\mu_1 = H|M) = 1$ *or* $\Pr(\mu_1 = H|B) = \frac{q(1-p)}{1-qp}$*.*

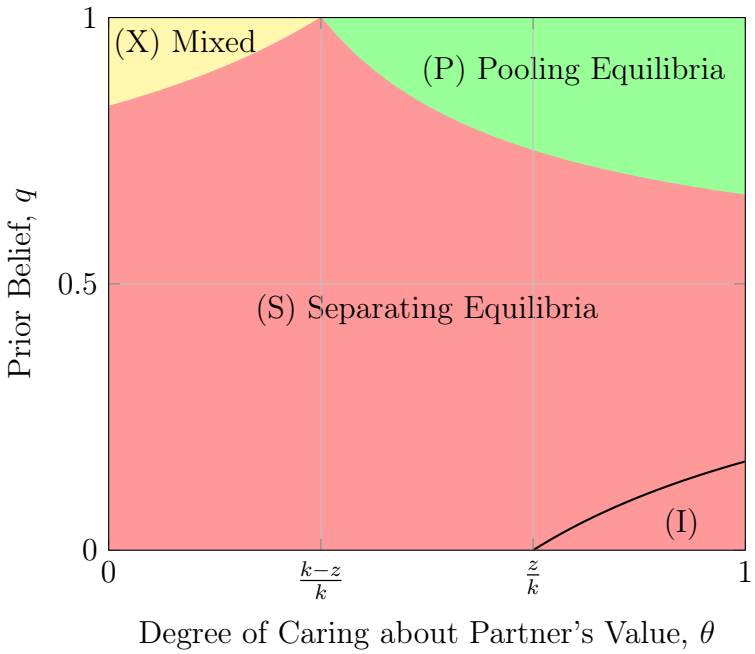
At $t = 3$*, a high-type player 1 accepts both* $\tau_H(\theta)$ *and* $\tau_L(\theta)$*, and a low-type player 1 accepts* $\tau_H(\theta)$ *but rejects* $\tau_L(\theta)$*. Player 1's posterior belief:* $\Pr(\mu_2 = H|\tau_H) = 1$ *or* $\Pr(\mu_2 = H|\tau_L) = \frac{q(1-r)}{1-qr}$*.*

The mixing probabilities are $p(\theta) = \frac{(2z+2k)q-2z+\theta(E_1[\mu_2|\tau_L]-k)q}{2kq+\theta(E_1[\mu_2|\tau_L]-k)q}$*, where* $E_1[\mu_2|\tau_L] = k\left(\frac{2q-qr(\theta)-1}{1-qr(\theta)}\right)$*, and* $r(\theta) = \frac{z+k+\theta k(2q-1)}{2kq}$*.*

In the mixed strategy equilibrium, a high-type player 1 sometimes initiates bargaining because there is gain if he meets a high-type player 2 who happens to give a positive payment $\tau_H(\theta) = -z + k - \theta k$. However, this gain is decreasing in θ, so the high-type player 1 also becomes less likely to initiate bargaining. As a result, the threshold $\hat{q}(\theta)$ approaches one as θ rises from zero to $\frac{k-z}{k}$. Consequently, the range of q where mixed strategy equilibria exist diminishes (see Area X in Figure 3.3) and the range for separating equilibria rises (see Area S in Figure 3.3 for $0 < \theta < \frac{k-z}{k}$). If $\theta = 1 - \frac{z}{k}$, there exists only separating equilibria for any level of q, the prior belief about the partner's type.

Now we turn to the case where players care about their partner's value of the marriage to even higher degrees $\frac{k-z}{k} \leq \theta \leq 1$. In addition to the separating equilibria, Figure 3.3 shows that there emerges a class of pooling equilibria in which both high and low types of player 1 choose to stay married (M) and refrain from initiating bargaining so that all married couples remain in their marriages (see Area P). In this equilibrium, even player 1 who has a low valuation of the marriage chooses M because of two reasons. First, they care sufficiently about their partner's value of the marriage. Second, they have sufficiently strong prior belief that their partner is a high type. As a result,

Figure 3.3: Equilibrium Analysis



such a low-type player 1 derives enough satisfaction from having a strong belief that their partner has a high value of the marriage to offset their own low value of the marriage. This occurs whenever

$$\begin{aligned} & u_{\langle 1,L\rangle}(M) \geq u_{<1,L>}(B) \\ \Rightarrow \quad & z - k + \theta k(2q-1) \geq 0 \quad \Rightarrow \quad q \geq \frac{-z+k+\theta k}{2\theta k}. \end{aligned}$$

Otherwise, the two types of player 1 choose different strategies in the separating equilibrium. The next two propositions describe these two classes of equilibria if players care sufficiently about their partner's value of the marriage.

Proposition 6. *If* $\frac{k-z}{k} \leq \theta \leq 1$ *and* $0 \leq q \leq \bar{q}(\theta)$*, there exists a separating equilibrium, where* $\bar{q}(\theta) \equiv \frac{-z+k+\theta k}{2\theta k}$*:*

> *At* $t = 1$*, a high-type player 1 stays married* (M) *and a low-type player 1 initiates bargaining* (B)*.*
>
> *At* $t = 2$*, a high-type player 2 demands* $\tau_H(\theta) = -z + k - \theta k$ *and a low-type player 2 demands* $\tau_L(\theta) = \min\{-z-k+\theta k, z-k-\theta k\}$*. Player 2's posterior belief:* $\Pr(\mu_1 = H|M) = 1$ *or* $\Pr(\mu_1 = H|B) = 0$*.*

At $t = 3$, a high-type player 1 always accepts $\tau_H(\theta)$ and only accepts $\tau_L(\theta)$ if $\theta \leq \frac{z}{k}$, and a low-type player 1 accepts $\tau_H(\theta)$ but rejects $\tau_L(\theta)$. Player 1's posterior belief: $\Pr(\mu_2 = H|\tau_H) = 1$ *or* $\Pr(\mu_2 = H|\tau_L) = 0$.

Proposition 7. *If $\frac{k-z}{k} \leq \theta \leq 1$ and $\bar{q}(\theta) \leq q \leq 1$, there exists a pooling equilibrium, where $\bar{q}(\theta) = \frac{-z+k+\theta k}{2\theta k}$.*

At $t = 1$, both high and low types of player 1 stay married (M).

At $t = 2$, a high-type player 2 demands $\tau_H(\theta) = -z + k - \theta k$ and a low-type player 2 demands $\tau_L(\theta) = \min\{-z-k+\theta k, z-k-\theta k\}$. Player 2's posterior belief: $\Pr(\mu_1 = H|M) = q$ *or* $\Pr(\mu_1 = H|B) = 0$.

At $t = 3$, a high-type player 1 always accepts $\tau_H(\theta)$ and only accepts $\tau_L(\theta)$ if $\theta \leq \frac{z}{k}$, and a low-type player 1 accepts $\tau_H(\theta)$ but rejects $\tau_L(\theta)$. Player 1's posterior belief: $\Pr(\mu_2 = H|\tau_H) = 1$ *or* $\Pr(\mu_2 = H|\tau_L) = 0$.

In the separating equilibrium, a high-type player 2 does not give but demands a negative transfer $\tau_H(\theta) = -z+k-\theta k$ from player 1 (if $\theta > \frac{k-z}{k}$). It may seem surprising that a low-type player 1 is still willing to accept $\tau_H(\theta)$ and meet this demand in equilibrium. The reason is that $\tau_H(\theta)$ is able to signal that player 2 is a high type (since a low-type player 2 chooses a different strategy) and this signal can raise player 1's marital utility significantly. Moreover, the impact of signalling one's high type on the partner's utility increases in θ. Therefore, as players care more about their partner's value of the marriage, a high-type player 2 is able to place an increasingly higher demand on a low-type player 1 in equilibrium.

A low-type player 2 chooses $\tau_L(\theta)$ as the minimum of two possible demands of negative transfers in equilibrium. One is $-z-k+\theta k$, the maximum amount that a high-type player 1 is willing to give to him or her, but this amount decreases in θ in absolute value. The other is $z-k-\theta k$, the minimum amount that he or she requires to be content with a marriage with a low type, but this amount increases in θ in absolute value. Thus, a higher θ, a greater concern for the partner's value, means that a high-type player 1 is less willing to give to a low type, and a low-type player 2 requires more from a low-type player 1. As a result, if $\theta > \frac{z}{k}$ both high and low types of player 1 would reject a demand of $\tau_L(\theta)$ from a low-type player 2.

One interesting phenomenon is that an inefficient marriage can be restored in the separating equilibrium through bargaining and signalling if $\theta > \frac{z}{k}$.

A marriage between a low type and a high type is initially inefficient if

$$\begin{aligned} & u_{\langle 1,L\rangle} + u_{\langle 2,H\rangle} < 0 \\ \Rightarrow\ & z - k + \theta k(2q-1) + z + k + \theta k(2q-1) < 0 \ \Rightarrow\ q < \frac{-z+\theta k}{2\theta k}. \end{aligned}$$

If players have a sufficiently low prior belief that their partner is a high type, their marriage is jointly sub-optimal. However, in equilibrium, a low-type player 1 would initiate bargaining, and a high-type player 2 would choose $\tau_H(\theta)$, which is accepted by player 1. The joint outcome in this equilibrium turns out to be an efficient marriage:

$$u_{\langle 1,L\rangle}(B, \tau_H(\theta), A) + u_{\langle 2,H\rangle}(B, \tau_H(\theta), A) = 2z > 0.$$

The reason is that the revelation of player 2 being a high type significantly raises player 1's low prior belief q and thus restores an initially sub-optimal marriage. This phenomenon occurs if θ is sufficiently high and q is low enough, as is shown in Area I in Figure 3.3.

3.4.2 Welfare Analysis

Now we examine the ex ante expected welfare of getting married for both having and not having a possibility to bargain with the threat of divorce. The assumption is that ex ante any player would have a high value of the marriage with probability q. When there is the possibility for divorce, it is ex ante equally likely to be the first or second player in the bargaining procedure.

The analysis focuses on the impact of θ, the degree of concern for the partner's value of the marriage, on the ex ante expected welfare of getting married with or without the possibility of divorce. As a benchmark, when there is no possibility of divorce, the ex ante expected welfare of getting married is

$$W_N(\theta) = q(z+k) + (1-q)(z-k) + \theta k(2q-1). \tag{3.3}$$

The first two terms are the expectation of one's own value of the marriage plus z and the last term is the expectation of the partner's value of the marriage multiplied by θ. Since players do not know their partner's true value, the expected welfare $W_N(\theta)$ might increase or decrease in θ depending on whether

q is above or below 1/2.

When players can bargain with divorce threat, there exists three classes of equilibria, as is shown in the previous section. For the class of separating equilibria characterized in Proposition 4 (if $0 \leq \theta < \frac{k-z}{k}$ and $0 < q < \hat{q}(\theta)$)and Proposition 6 (if $\frac{k-z}{k} \leq \theta \leq 1$ and $0 < q < \bar{q}(\theta)$), the ex ante expected welfare of getting married is

$$W_S(\theta) = q(z+k) + q(1-q)(z-k) + \theta k q^2. \tag{3.4}$$

The first two terms of $W_S(\theta)$ are the same as the first two terms of $W_N(\theta)$ in equation (3.3) minus $(1-q)^2(z-k)$, which is the negative expected payoff of being in a marriage of two low types. The kind of couples always ends up in divorce in the separating equilibrium. The last term is the expectation of the partner's value of the marriage, which might be revealed through the process of bargaining.

Note that $W_S(\theta)$ is increasing in θ for any $q > 0$ in the separating equilibrium. This means that a greater degree of concern for the partner's value of the marriage increases the ex ante expected welfare of getting married with the possibility to bargain with divorce threat in the separating equilibrium.

For the mixed strategy equilibria characterized in Proposition 5 (if $0 \leq \theta < \frac{k-z}{k}$ and $\hat{q}(\theta) < q < 1$), the ex ante expected welfare is

$$W_X(\theta) = q(z+k) + q(1-q)(z-k) \underbrace{-(1-q)q(1-r(\theta))z}_{\text{L meeting H who chooses } \tau_L(\theta)} + \theta k q^2. \tag{3.5}$$

The third term of $W_X(\theta)$ is the expected loss when a low-type player 1 meets a high-type player 2 who happens to demand $\tau_L(\theta)$, which is rejected by player 1. The other terms are the same as $W_S(\theta)$ in the separating equilibrium.

Even though there is inefficient divorce in the mixed strategy equilibrium, $W_X(\theta)$ is still increasing in θ. The reason is that as players care more about their partner's value of the marriage, the incentive for a high-type player 1 to choose a mixed strategy to bargain inefficiently falls. Thus, the expected welfare of getting married increases even more when there exists mixed strategy equilibria.

For the pooling equilibria characterized in Proposition 7 (if $\frac{k-z}{k} \leq \theta \leq 1$ and $\bar{q}(\theta) \leq q \leq 1$), the ex ante expected welfare of getting married is exactly the same as $W_N(\theta)$ in equation (3.3). The reason is that both high and low

types of player 1 choose to stay married without using the divorce threat to bargain . In this equilibrium, refraining from initiating bargaining can no longer signal one's high value of the marriage because both types choose this same strategy.

Therefore, a higher θ or concern for the partner's value of the marriage improves the ex ante expected welfare of getting married in both separating equilibrium ($W_S(\theta)$) and mixed strategy equilibrium ($W_X(\theta)$), but not necessarily in pooling equilibrium or when there is no possibility of divorce.

3.5 Conclusion

This chapter explores the welfare consequences of having a possibility to bargain with divorce threat. To do this, we develop a bargaining model with two-sided asymmetric information about the partner's value of the marriage. We also examine the impact of incorporating interdependent preferences on the expected welfare of getting married when there is the possibility of divorce.

Our notion of interdependent preference is that spouses care about their partner's hidden value of the marriage. This feature causes signalling one's type to have a direct impact on the partner's utility. If a spouse is revealed to have a high value, the partner becomes more satisfied with the marriage and thus more tolerant of the first spouse's action, even if the action is unfavourable. This demonstrates the significant impact of signalling one's hidden value on each other's interdependent utility.

The model shows that some jointly optimal marriages can end up in divorce if spouses do not care whether their partner has a high or low value of the marriage, yet believing that their partner has a high value and is willing to give a positive payment. As a result, spouses try to extract too much rent from each other, leading efficient marriages to divorce.

The main result is that incorporating interdependent preferences can eliminate the possibility of inefficient divorce generated by asymmetric information. A greater concern for the partner's value of the marriage generally improves and never reduces the ex ante welfare of getting married when there is the possibility to divorce. This increased welfare is gained by the opportunity to signal each other's private value of the marriage. Sometimes even an initially inefficient marriage can be turned into a jointly optimal marriage after bargaining and signalling one's high value of the marriage.

In summary, if spouses care sufficiently much about their partner's value of the marriage, having the possibility to divorce yields an ex ante welfare that is at least as high as having no possibility to divorce. In contrast, if spouses care little about their partner's value and hold a sufficiently strong prior belief that their partner has a high value of the marriage, having the possibility to divorce can yield lower ex ante welfare than having no possibility to divorce.

3.6 Appendix

Proof of Proposition 1. The specified strategies constitute a separating equilibrium because no player has an incentive to unilaterally deviate to another strategy. At $t = 3$, a high-type player 1 accepts a proposed transfer if and only if the marital utility is higher than the divorce outcome:

$$u_{\langle 1,H\rangle}(B,\tau,A) = z + H + \tau \geq 0 \;\Rightarrow\; \tau \geq -z - H. \tag{3.6}$$

Similarly, a low-type player 1 accepts a transfer if and only if

$$u_{\langle 1,L\rangle}(B,\tau,A) = z - L + \tau \geq 0 \;\Rightarrow\; \tau \geq -z + L. \tag{3.7}$$

At $t = 2$, in the separating equilibrium the second players know that player 1 who initiated bargaining must be a low type and that a low-type player 1 rejects any proposal less than $-z + L$. If a low-type player 2 deviates to choose giving $\tau_H = -z + L$, he would end up with a negative payoff of $2z - 2L$, which is worse than the divorce outcome that he gets from demanding a payment $\tau_L = -z - H$. If a high-type player 2 deviates to choose demanding τ_L, he would be rejected and receive the divorce outcome, which is worse than giving a transfer τ_H that is accepted by a low-type player 1 and ending up with positive marital utility of $2z + H - L$. Therefore, neither type of player 2 would deviate from their strategies in the equilibrium.

At $t = 1$, a low-type player 1 would never choose M because he gets a negative marital payoff of $z - L$, which is strictly less than the divorce outcome. A high-type player 1 would deviate to choose B only if the expected utility of initiating bargaining is higher than the utility of staying married. Thus, if condition (3.1) is satisfied, a high-type player 1 will always choose M and there exists a separating equilibrium. □

Proof of Proposition 2. At $t = 3$, the same individual rationality conditions (3.6) and (3.7) apply in the mixed strategy equilibrium. In this equilibrium, a high-type player 1 must be indifferent between choosing M and B at $t = 1$:

$$\begin{aligned} & E[u_{\langle 1,H\rangle}(M)|r] = E[u_{\langle 1,H\rangle}(B)|r] \\ \Rightarrow \quad & z + H = qr(H + L) \quad \Rightarrow \quad r = \frac{z + H}{q(H + L)}. \end{aligned}$$

For r to be less than one, q must be greater than $\frac{z+H}{L+H}$. After observing player 1 choosing B, player 2 updates the posterior probability that player 1 has a high valuation of the marriage:

$$\Pr(\mu_1 = H|B) = \frac{q(1-p)}{q(1-p) + (1-q)}.$$

In the equilibrium, a high-type player 2 must also be indifferent between choosing $\tau_H = -z + L$ and $\tau_L = -z - H$ at $t = 2$:

$$\begin{aligned} & E[u_{\langle 2,H\rangle}(B, \tau_H)|p] = E[u_{\langle 2,H\rangle}(B, \tau_L)|p] \\ \Rightarrow \quad & 2z + H - L = \Pr(\mu_1 = H|B) * (2z + 2H) \\ \Rightarrow \quad & p = \frac{q(2z + 2H) - (2z + H - L)}{q(H + L)} \end{aligned}$$

For p less than one, q must be less than one. For p to be greater than one, q must be greater than $\frac{2z+H-L}{2z+2H}$, which is less than $\frac{z+H}{L+H}$. Thus, there exists a mixed strategy equilibrium only if $\frac{z+H}{L+H} < q < 1$. □

Proof of Claim 1. The only possible pooling equilibrium is when $a_{\langle 1,H\rangle} = a_{\langle 1,L\rangle} = B$. Suppose there exists such a pooling equilibrium. Then player 2's posterior belief is $\Pr(\mu_1 = H|B) = q$. A low-type player 2 always chooses $\tau_L = -z - H$, but a high-type player 2 may choose τ_L or $\tau_H = -z + L$. If a high-type player 2 chooses τ_L, a high-type player 1 will always deviate to choose M because $u_{\langle 1,H\rangle}(M) > u_{\langle 1,H\rangle}(B) \Rightarrow z + H > 0$. This contradicts with the existence of the pooling equilibrium. If a high-type player 2 chooses τ_H, the incentive compatibility requires that $E(u_{\langle 2,H\rangle}(B, \tau_H)) \geq E(u_{\langle 2,H\rangle}(B, \tau_L))$

$$\Rightarrow \quad q \leq \frac{2z + H - L}{2z + 2H}. \tag{3.8}$$

By Assumption 1, condition (3.8) implies that condition (3.1) holds, suggesting

that a high-type player 1 will deviate to choose M. This again contradicts with the existence of the pooling equilibrium. □

Proof of Proposition 3. If $0 \leq q \leq \hat{q}$, $W_S - W_N = -(1-q)^2(z-L) > 0$. If $\hat{q} < q \leq 1$, $W_X - W_N = -(1-q)^2(z-L) - \left(\frac{1}{2}\right)[(1-q)q(1-r)(2z+H-L)]$. $W_X - W_N < 0$ if and only if $\left(1 - \frac{(L-z)(2z+H-L)}{(H+L)^2}\right) < q < 1$.

□

Proof of Propositions 4. This proposition is a generalization of Proposition 1, so the proof is very similar. At $t = 3$, a high-type player 1 accepts a proposed transfer if and only if the marital utility is higher than the divorce outcome:

$$u_{\langle 1,H\rangle}(B, \tau, A) \geq 0 \;\Rightarrow\; \tau \geq -z - k - \theta E_1[\mu_2|\tau]. \tag{3.9}$$

Similarly, a low-type player 1 accepts a transfer if and only if

$$u_{\langle 1,L\rangle}(B, \tau, A) \geq 0 \;\Rightarrow\; \tau \geq -z + k - \theta E_1[\mu_2|\tau]. \tag{3.10}$$

In the separating equilibrium player 2 knows that $\Pr(\mu_1 = H|B) = 0$ and player 1 knows that $\Pr(\mu_2 = H|\tau_H) = 1$ and $\Pr(\mu_2 = H|\tau_L) = 0$. If a low-type player 2 deviates to giving $\tau_H = -z + k - \theta k$, he will end up with a lower payoff in marriage. If a high-type player 2 deviates to demanding $\tau_L = -z - k + \theta k$, he will end up with a divorce outcome because a low-type player 1 will reject τ_L. This zero outcome is less than a positive outcome of $2z$ by giving τ_H that will be accepted by a low-type player 1. Therefore, neither type of player 2 would deviate from their strategies in the equilibrium.

At $t = 1$, a low-type player 1 would deviate to choose M if $u_{\langle 1,L\rangle}(M) \geq u_{\langle 1,L\rangle}(B) \;\Rightarrow\; z - k + \theta k(2q-1) \geq 0$

$$\Rightarrow \quad q \geq \frac{-z + k + \theta k}{2\theta k} = \bar{q}(\theta). \tag{3.11}$$

Since the range of θ for this equilibrium is $0 \leq \theta < \frac{k-z}{k}$, $\bar{q}(\theta)$ is always greater than one. Thus, a low-type player 1 will not deviate to choose M. A high-type player 1 will deviate to choose B if $E(u_{\langle 1,H\rangle}(B)) \geq E(u_{\langle 1,H\rangle}(M)) \;\Rightarrow\; q(2k) \geq z + k + \theta k(2q-1)$

$$\Rightarrow\; q \geq \frac{z + k - \theta k}{2k - 2\theta k} = \hat{q}(\theta) \tag{3.12}$$

Thus, given that $q \leq \hat{q}(\theta)$, a high-type player 1 will choose M and there exists

a separating equilibrium. □

Proof of Proposition 5. At $t = 3$, the individual rationality conditions (3.9) and (3.10) both apply in the mixed strategy equilibrium. In this equilibrium, a high-type player 1 must be indifferent between choosing M and B at $t = 1$:

$$\begin{aligned} & E[u_{\langle 1,H\rangle}(M)|r(\theta)] = E[u_{\langle 1,H\rangle}(B)|r(\theta)] \\ \Rightarrow \quad & z + k + \theta k(2q-1) = qr(2k) \quad \Rightarrow \quad r(\theta) = \frac{z + k + \theta k(2q-1)}{2kq}. \end{aligned}$$

For $r(\theta)$ to be less than one, q must be greater than $\hat{q}(\theta)$. After observing player 1 choosing B, player 2 updates the posterior probability that player 1 has a high valuation of the marriage:

$$\Pr(\mu_1 = H|B) = \frac{q(1-p)}{q(1-p) + (1-q)}.$$

In the equilibrium, a high-type player 2 must also be indifferent between choosing $\tau_H(\theta) = -z + k - \theta k$ and $\tau_L(\theta) = -z - k - \theta E_1[\mu_2|\tau_L]$ at $t = 2$:

$$\begin{aligned} & E[u_{\langle 2,H\rangle}(B,\tau_H)|p(\theta)] = E[u_{\langle 2,H\rangle}(B,\tau_L)|p(\theta)] \\ \Rightarrow \quad & 2z + \theta(k + E_1[\mu_1|B]) = \Pr(\mu_1 = H|B)(2z + 2k + \theta(k + E_1[\mu_2|\tau_L])) \\ \Rightarrow \quad & p(\theta) = \frac{(2z+2k)q - 2z + \theta(E_1[\mu_2|\tau_L] - k)q}{2kq + \theta(E_1[\mu_2|\tau_L] - k)q}, \end{aligned}$$

where $E_1[\mu_2|\tau_L] = k\left(\frac{2q-qr-1}{1-qr}\right)$. A low-type player 2 chooses the same $\tau_L(\theta)$ because any $\tau > \tau_L(\theta)$ yields a lower payoff and any $\tau < \tau_L(\theta)$ is rejected by player 1. Thus, no player has an incentive to deviate from the specified mixed strategy equilibrium if $\hat{q}(\theta) < q < 1$. □

Proof of Proposition 6. The separating equilibria in the range of $\frac{k-z}{k} \leq \theta \leq 1$ are similar to the separating equilibria in the range $0 \leq \theta < \frac{k-z}{k}$ (in Proposition 4) except a few differences. First, according to individual rationality condition (3.9), a high-type player 1 at $t = 3$ accepts $\tau_L(\theta)$ if $\theta \leq \frac{z}{k}$ and rejects it if $\theta > \frac{z}{k}$. Second, a low-type player 2 chooses $\tau_L(\theta) = \min\{-z - k + \theta k, z - k - \theta k\}$, which is always rejected by a low-type player 1. Again, any deviation from this strategy either yields a lower payoff or is still rejected. Third, according to the constraint (3.11), q must be less than or equal to $\bar{q}(\theta)$ so that a low-type player 1 will not deviate to choose M. □

Proof of Proposition 7. The pooling equilibrium in which both types of player 1 choose M exists if the constraint (3.11) is satisfied. The game then ends at $t = 1$, and the second and third stages of the bargaining procedure is not reached. Since any belief off the equilibrium path is consistent with the other player's strategy, we specify off-equilibrium beliefs and strategies to be the same as those characterized in the separating equilibrium in Proposition 6. □

Bibliography

Akerlof, Robert. 2015. Anger and Enforcement. Unpublished Manuscript, University of Warwick.

Alós-Ferrer, Carlos, & Hügelschäfer, Sabine. 2012. Faith in intuition and behavioral biases. *Journal of Economic Behavior & Organization*, **84**(1), 182–192.

Anderson, Christopher M, & Putterman, Louis. 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, **54**(1), 1–24.

Andreoni, James, & Miller, John. 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, **70**(2), 737–753.

Basu, Kaushik. 2006. Gender and Say: a Model of Household Behaviour with Endogenously Determined Balance of Power. *Economic Journal*, **116**(511), 558–580.

Becker, Gary. 1991. *Treatise on the Family,*. Harvard University Press.

Betsch, Cornelia, & Kunz, Justus J. 2008. Individual strategy preferences and decisional fit. *Journal of Behavioral Decision Making*, **21**(5), 532–555.

Bewley, Truman F. 1999. *Why wages don't fall during a recession*. Harvard University Press.

Björklund, Fredrik, & Bäckström, Martin. 2008. Individual differences in processing styles: validity of the Rational–Experiential Inventory. *Scandinavian Journal of Psychology*, **49**(5), 439–446.

Bolton, Gary E, & Ockenfels, Axel. 2000. ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, **90**(1), 166–193.

Brandts, Jordi, & Charness, Gary. 2011. The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, **14**(3), 375–398.

Browning, Martin. 2009. Love, Betrayal and Commitment. University of Oxford.

Browning, Martin, & Chiappori, Pierre-André. 1998. Efficient Intra-Household Allocations: A General Characterization and Empirical Tests. *Econometrica*, **66**(6), 1241–1278.

Browning, Martin, Chiappori, Pierre-André, & Lechene, Valérie. 2010. Distributional effects in household models: separate spheres and income pooling. *Economic Journal*, **120**(545), 786–799.

Browning, Martin, Chiappori, Pierre-André, & Weiss, Yoram. 2014. *Family Economics*. Oxford University Press.

Cacioppo, John T, & Petty, Richard E. 1982. The need for cognition. *Journal of Personality and Social Psychology*, **42**(1), 116.

Cacioppo, John T, Petty, Richard E, Feinstein, Jeffrey A, & Jarvis, W Blair G. 1996. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, **119**(2), 197.

Camerer, Colin, & Thaler, Richard H. 1995. Anomalies: Ultimatums, dictators and manners. *Journal of Economic Perspectives*, **9**(2), 209–219.

Camerer, Colin F, & Hogarth, Robin M. 1999. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty*, **19**(1-3), 7–42.

Carlsmith, Kevin M, Darley, John M, & Robinson, Paul H. 2002. Why do we punish?: Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, **83**(2), 284.

Carnevale, Jessica J, Inbar, Yoel, & Lerner, Jennifer S. 2011. Individual differences in need for cognition and decision-making competence among leaders. *Personality and Individual Differences*, **51**(3), 274–278.

Carpenter, Jeffrey, & Matthews, Peter Hans. 2009. What norms trigger punishment? *Experimental Economics*, **12**(3), 272–288.

Carpenter, Jeffrey P. 2007. The demand for punishment. *Journal of Economic Behavior & Organization*, **62**(4), 522–542.

Charness, Gary, & Rabin, Matthew. 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics*, **117**(3), 817–869.

Charness, Gary, Gneezy, Uri, & Kuhn, Michael A. 2012. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, **81**(1), 1–8.

Chiappori, Pierre-André. 1988. Rational Household Labor Supply. *Econometrica*, **56**(1), 63–90.

Chiappori, Pierre-André, Iyigun, Murat, & Weiss, Yoram. 2007. Public Goods, Transferable Utility and Divorce Laws. IZA Working Paper No. 2646.

Cohen, Arthur R, Stotland, Ezra, & Wolfe, Donald M. 1955. An experimental investigation of need for cognition. *Journal of Abnormal and Social Psychology*, **51**(2), 291.

Cooter, Robert. 1998. Expressive Law and Economics. *Journal of Legal Studies*, **27**(2), 585–608.

Danziger, Shai, Moran, Simone, & Rafaely, Vered. 2006. The influence of ease of retrieval on judgment as a function of attention to subjective experience. *Journal of Consumer Psychology*, **16**(2), 191–195.

Dewatripont, Mathias. 1989. Renegotiation and Information Revelation Over Time: The Case of Optimal Labor Contracts. *Quarterly Journal of Economics*, **104**, 589–619.

Dewatripont, Mathias, & Maskin, Eric. 1990. Contract Renegotiation in Models of Asymmetric Information. *European Economic Review*, **34**, 311–321.

Dufwenberg, Martin, & Kirchsteiger, Georg. 2004. A theory of sequential reciprocity. *Games and Economic Behavior*, **47**(2), 268–298.

Durham, Alexis M. 1987. Justice in Sentencing: The Role of Prior Record of Criminal Involvement. *Journal of Criminal Law and Criminology*, **78**(3), 614–643.

Ellingsen, Tore, & Johannesson, Magnus. 2008. Pride and Prejudice: The Human Side of Incentive Theory. *American Economic Review*, **98**(3), 990–1008.

Epstein, Seymour. 2010. Demystifying intuition: What it is, what it does, and how it does it. *Psychological Inquiry*, **21**(4), 295–312.

Epstein, Seymour. 2014. *Cognitive-experiential theory: An integrative theory of personality*. Oxford University Press.

Epstein, Seymour, & Pacini, Rosemary. 1999. Some basic issues regarding dual-process theories from the perspective of cognitive-experiential self-theory. *Dual-process Theories in Social Psychology*, 462–482.

Epstein, Seymour, Pacini, Rosemary, Denes-Raj, Veronika, & Heier, Harriet. 1996. Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, **71**(2), 390.

Evans, Jonathan St BT, & Stanovich, Keith E. 2013. Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, **8**(3), 223–241.

Falk, Armin, & Fischbacher, Urs. 2006. A theory of reciprocity. *Games and Economic Behavior*, **54**(2), 293–315.

Falk, Armin, Fehr, Ernst, & Fischbacher, Urs. 2005. Driving Forces behind Informal Sanctions. *Econometrica*, **73**(6), 2017–2030.

Falk, Armin, Fehr, Ernst, & Fischbacher, Urs. 2008. Testing theories of fairnessâĂŤIntentions matter. *Games and Economic Behavior*, **62**(1), 287–303.

Fehr, Ernst, & Gächter, Simon. 2000a. Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, **90**(4), 980–994.

Fehr, Ernst, & Gächter, Simon. 2000b. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, **14**(3), 159–181.

Fehr, Ernst, & Gächter, Simon. 2002. Altruistic punishment in humans. *Nature*, **415**(6868), 137–140.

Fehr, Ernst, & Schmidt, Klaus M. 1999. A theory of fairness, competition, and cooperation. *Quarterly journal of Economics*, 817–868.

Fischbacher, Urs. 2007. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, **10**(2), 171–178.

Friedberg, Leora. 1998. Did unilateral Divorce Raise Divorce Rates? Evidence from Panel Data. *American Economic Review*, **88**(3), 608–627.

Friedberg, Leora, & Stern, Steven. 2014. Marriage, Divorce, and Asymmetric Information. *International Economic Review*, **55**(4), 1155–1199.

Gächter, Simon, & Renner, Elke. 2010. The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, **13**(3), 364–377.

Glaser, Markus, & Walther, Torsten. 2014. Run, Walk, or Buy? Financial Literacy, Dual-Process Theory, and Investment Behavior. Unpublished Manuscript.

Graham, Jesse, Haidt, Jonathan, Koleva, Sena, Motyl, Matt, Iyer, Ravi, Wojcik, Sean P, & Ditto, Peter H. 2012. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology*, **47**, 55–130.

Grimm, Veronika, & Mengel, Friederike. 2011. Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters*, **111**(2), 113–115.

Grossman, Stanford, & Hart, Oliver. 1986. The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration. *Journal of Political Economy*, **94**(4), 691–719.

Güth, Werner, Huck, Steffen, & Ockenfels, Peter. 1996. Two-level ultimatum bargaining with incomplete information: An experimental study. *The Economic Journal*, 593–604.

Haidt, Jonathan. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, **108**(4), 814.

Haidt, Jonathan, & Joseph, Craig. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, **133**(4), 55–66.

Halla, Martin, & Scharler, Johann. 2012. Marriage, Divorce and Interstate Risk Sharing. *Scandinavian Journal of Economics*, **114**(1), 55–78.

Hart, Oliver, & Moore, John. 1988. Incomplete Contracts and Renegotiation. *Econometrica*, **56**(4), 755–85.

Hart, Oliver, & Moore, John. 2008. Contracts as Reference Points. *Quarterly Journal of Economics*, **123**(1), 1–48.

Heidhues, Paul, & Kőszegi, Botond. 2008. Competition and price variation when consumers are loss averse. *American Economic Review*, **98**(4), 1245–1268.

Herrmann, Benedikt, Thöni, Christian, & Gächter, Simon. 2008. Antisocial punishment across societies. *Science*, **319**(5868), 1362–1367.

Hodgkinson, Gerard P, & Clarke, Ian. 2007. Conceptual note: Exploring the cognitive significance of organizational strategizing: A dual-process framework and research agenda. *Human Relations*, **60**(1), 243–255.

Hodgkinson, Gerard P, Sadler-Smith, Eugene, Sinclair, Marta, & Ashkanasy, Neal M. 2009. More than meets the eye? Intuition and analysis revisited. *Personality and Individual Differences*, **47**(4), 342–346.

Hopfensitz, Astrid, & Reuben, Ernesto. 2009. The importance of emotions for the effectiveness of social punishment. *Economic Journal*, **119**(540), 1534–1559.

Horton, John J, Rand, David G, & Zeckhauser, Richard J. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, **14**(3), 399–425.

Kahneman, Daniel. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kahneman, Daniel, & Tversky, Amos. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, **47**(2), 263–292.

Kahneman, Daniel, Knetsch, Jack L, & Thaler, Richard. 1986. Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, **76**(4), 728–741.

Kant, Immanuel. 1790. *The Science of Right*. Translated by W. Hastie.

Levin, Irwin P, Huneke, Mary E, & Jasper, John D. 2000. Information processing at successive stages of decision making: Need for cognition and inclusion–exclusion effects. *Organizational Behavior and Human Decision Processes*, **82**(2), 171–193.

Levine, David K. 1998. Modeling altruism and spitefulness in experiments. *Review of economic dynamics*, **1**(3), 593–622.

LoBue, Vanessa, Nishida, Tracy, Chiong, Cynthia, DeLoache, Judy S, & Haidt, Jonathan. 2011. When getting something good is bad: Even three-year-olds react to inequality. *Social Development*, **20**(1), 154–170.

Lundberg, Shelly, & Pollak, Robert. 1993. Separate Spheres Bargaining and the Marriage Market. *Journal of Political Economy*, **101**(6), 988–1009.

Mahoney, Kevin T, Buboltz, Walter, Levin, Irwin P, Doverspike, Dennis, & Svyantek, Daniel J. 2011. Individual differences in a within-subjects risky-choice framing study. *Personality and Individual Differences*, **51**(3), 248–257.

Mason, Winter, & Watts, Duncan J. 2010. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, **11**(2), 100–108.

Matouschek, Niko, & Rasul, Imran. 2008. The Economics of the Marriage Contract: Theories and Evidence. *Journal of Law and Economics*, **51**(1), 59–110.

Mazzocco, Maurizio. 2007. Household intertemporal behaviour: A collective characterization and a test of commitment. *Review of Economic Studies*, **74**(3), 857–895.

McElroy, Marjorie, & Horney, Mary. 1981. Nash-Bargained Household Decisions: Toward a Generalization of the Theory of Demand. *International Economic Review*, **22**(2), 333–49.

Myerson, Roger, & Satterthwaite, Mark. 1983. Efficient Mechanisms for Bilateral Trading. *Journal of Economic Theory*, **29**(2), 265–281.

Myrseth, Kristian Ove R, & Wollbrant, Conny E. 2015. *Intuitive cooperation refuted: Commentary on Rand et al.(2012) and Rand et al.(2014)*. Tech. rept. University of Gothenburg, Department of Economics.

Norenzayan, Ara, Smith, Edward E, Kim, Beom Jun, & Nisbett, Richard E. 2002. Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, **26**(5), 653–684.

Ostrom, Elinor, Walker, James, & Gardner, Roy. 1992. Covenants with and without a Sword: Self-governance Is Possible. *American Political Science Review*, **86**(02), 404–417.

Pacini, Rosemary, & Epstein, Seymour. 1999. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, **76**(6), 972.

Peters, Elizabeth. 1986. Marriage and divorce: informal constraints and private contracting. *American Economic Review*, **76**(3), 437–âĂŞ454.

Petty, Richard, & Cacioppo, John T. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.

Petty, Richard E, DeMarree, Kenneth G, Briñol, Pablo, Horcajo, Javier, & Strathman, Alan J. 2008. Need for cognition can magnify or attenuate priming effects in social judgment. *Personality and Social Psychology Bulletin*, **34**(7), 900–912.

Polinsky, A Mitchell, & Rubinfeld, Daniel L. 1991. A model of optimal fines for repeat offenders. *Journal of Public Economics*, **46**(3), 291–306.

Rabin, Matthew. 1993. Incorporating fairness into game theory and economics. *American economic review*, **83**(5), 1281–1302.

Rainer, Helmut. 2007. Should we write prenuptial contracts? *European Economic Review*, **51**(2), 337–363.

Rand, David G, Greene, Joshua D, & Nowak, Martin A. 2012. Spontaneous giving and calculated greed. *Nature*, **489**(7416), 427–430.

Rand, David G, Peysakhovich, Alexander, Kraft-Todd, Gordon T, Newman, George E, Wurzbacher, Owen, Nowak, Martin A, & Greene, Joshua D. 2014. Social heuristics shape intuitive cooperation. *Nature Communications*, **5**(3677).

Rasul, Imran. 2006. The Economics of Child Custody. *Economica*, **73**(289), 1–25.

Roberts, Julian V. 1997. The role of criminal record in the sentencing process. *Crime and Justice*, **22**, 303–363.

Rubinstein, Ariel. 2007. Instinctive and cognitive reasoning: A study of response times. *Economic Journal*, **117**(523), 1243–1259.

Shiloh, Shoshana, Salton, Efrat, & Sharabi, Dana. 2002. Individual differences in rational and intuitive thinking styles as predictors of heuristic responses and framing effects. *Personality and Individual Differences*, **32**(3), 415–429.

Spier, Kathryn. 1992. Incomplete Contracts and Signalling. *RAND Journal of Economics*, **23**(3), 432.

Stevenson, Betsey. 2007. The Impact of Divorce Laws on Marriage-Specific Capital. *Journal of Labor Economics*, **25**(1), 75–94.

Stevenson, Betsey, & Wolfers, Justin. 2006. Bargaining in the Shadow of the Law: Divorce Laws and Family Distress. *Quarterly Journal of Economics*, **121**(1), 267–288.

Sutter, Matthias, Kocher, Martin, & Strauß, Sabine. 2003. Bargaining under time pressure in an experimental ultimatum game. *Economics Letters*, **81**(3), 341–347.

Sylwester, Karolina, Herrmann, Benedikt, & Bryson, Joanna J. 2013. Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, **6**(3), 167.

Tabibnia, Golnaz, Satpute, Ajay B, & Lieberman, Matthew D. 2008. The sunny side of fairness preference for fairness activates reward circuitry (and

disregarding unfairness activates self-control circuitry). *Psychological Science*, **19**(4), 339–347.

Tinghög, Gustav, Andersson, David, Bonn, Caroline, Böttiger, Harald, Josephson, Camilla, Lundgren, Gustaf, Västfjäll, Daniel, Kirchler, Michael, & Johannesson, Magnus. 2013. Intuition and cooperation reconsidered. *Nature*, **498**(7452), E1–E2.

Trent, Jason, & King, Laura A. 2013. Faith in Intuition moderates the effects of positive affect on gender stereotyping. *Personality and Individual Differences*, **54**(7), 865–868.

Trounstine, Jean. 2014. Brutal Crimes Don't Justify Bad Laws. *Truthout*. 26 October.

Tversky, Amos, & Kahneman, Daniel. 1981. The framing of decisions and the psychology of choice. *Science*, **211**(4481), 453–458.

Verkoeijen, Peter PJL, & Bouwmeester, Samantha. 2014. Does Intuition Cause Cooperation? *PLoS ONE*, **9**(5).

Wickelgren, Abraham L. 2009. Why Divorce Laws Matter: Incentives for Noncontractible Marital Investments under Unilateral and Consent Divorce. *Journal of Law, Economics, and Organization*, **25**(1), 80–106.

Witteman, Cilia, van den Bercken, John, Claes, Laurence, & Godoy, Antonio. 2009. Assessing rational and intuitive thinking styles. *European Journal of Psychological Assessment*, **25**(1), 39–47.

Wolfers, Justin. 2006. Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results. *American Economic Review*, **96**(5), 1802–1820.

Zhylyevskyy, Oleksandr. 2012. Spousal Conflict and Divorce. *Journal of Labor Economics*, **30**(4), 915–962.