Essays on Behavioural Game Theory

Sigbjørn Birkeland d.y.

Bergen, April 2011

Contents

Acknowledgements

I am grateful to the Norwegian School of Economics and Business Administration for the financial support and opportunity to undertake research in economics. It would not have been such a great experience without the advice and inspiration of Bertil Tungodden. He awoke my interest as a student in the fundamental questions of economic analysis and political philosophy. I am very thankful for his advice on the research questions and his detailed comments on the following essays.

I would very much like to thank James Konow and Alexander W. Cappelen for comments on the essays and helpful discussions on many of the topics of this thesis. Being part of an excellent research group in experimental economics has been of great inspiration and a learning experience, and I would like to thank Erik Ø. Sørensen, Ingvild Almås, and Knut Nygaard.

In addition, I would like to thank Victor D. Norman, who gave me the opportunity to be a research assistant in economics studying for the Høyere Avdeling. This thesis was written when I was on leave from Storebrand ASA, and I am grateful for their support.

This work would not have been possible without the love and encouragement of my wife Anne-Lise, and our boys, Hans, Petter, and Fredrik.

Introduction: Behavioural game theory

Game theory is the standard method in economics used to analyse situations where people or firms interact, for example, auctions, bargaining, cooperation, markets with a small number of firms, and many other social dilemmas such as the provision of public goods. Behavioural game theory is an approach to economics that uses psychological regularities to suggest ways to weaken rationality assumptions and extend the motivational basis for economic behaviour (Camerer, 2003). The three following essays are all contributions to behavioural game theory.

The first two essays are about bargaining, one of the basic activities of economic life. Bargaining is the process by which parties agree on the terms of a transaction, the distribution of costs and gains, and settle disputes. Many bargaining processes are inefficient and result in agreements where both parties could have achieved a better outcome (Johansen, 1979). There are a number of different mechanisms that can improve efficiency in bargaining, of which one of the most common is to include the option to let a third party decide on the issue. The first essay is about how a third party mechanism influences bargaining. It shows that, under reasonable conditions, the possibility of a third party decision will improve efficiency in bargaining. This contrasts with the established hypothesis of 'the chilling effect', where a possible third party decision reduces efficiency in bargaining because a compromising third party motivates the bargainers to stick to extreme positions during the bargaining process (Stevens, 1966).

There are many possible outcomes of bargaining, and a number of theories have been developed to understand which outcomes will be reached under different circumstances. The second essay in the thesis is about how individuals motivated by fairness considerations affect the bargaining outcome. Recent research on fairness in bargaining has developed models that include fairness but, in almost all cases, the analysis is limited to players agreeing on the same fairness principle (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). In the second essay, I extend the analysis to situations where individuals follow different fairness principles. This model is used to analyse the influence of fairness motivation on the possibility of reaching an agreement in bargaining, and to examine the properties of the agreement. I show that bargaining between two individuals who are strongly fairness motivated, but who disagree about what represents a fair division, ends in disagreement. This result contrasts the standard bargaining model with individuals who are only motivated by material self-interest, which always leads to agreement. Furthermore, by applying the Nash bargaining solution, I study the influence of fairness motivation on the bargaining outcome. A fairness motivated individual reaches an outcome that is closer to his fairness principle in bargaining against an individual who is only motivated by material self-interest.

In the third essay, a trust game is used to study the social preferences of prisoners and a benchmark group recruited from a representative sample of the

Norwegian population. Economists have traditionally focused on how economic opportunities affect criminal behaviour and have largely ignored the role of social preferences (Becker, 1968; Allingham and Sandmo, 1972; Andvig and Moene, 1990). The third paper studies the social preferences of criminals and it is, to our knowledge, the first to do so by conducting an economic experiment on a group of prisoners. Comparing the behaviour of the prisoners with the behaviour of a benchmark group, we find a striking similarity in the importance the two groups attach to pro-social preferences both in strategic and non-strategic situations. The pro-social behaviour of the prisoners in our experiment clearly contrasts with their anti-social behaviour outside the lab. One possible explanation for this cross-situational inconsistency is that behaviour in the lab is motivated by different social preferences than behaviour outside the lab. The situational inconsistency in behaviour could, however, also be seen as suggesting that social preferences are of little importance, compared to circumstances, in explaining criminal behaviour.

This introductory essay provides some background to the bargaining problem, the behavioural assumptions used in the modelling, and the experimental method that is used in the three following essays. These introductory remarks are meant to both describe selected literature within the research area, and to briefly discuss some methodological aspects of the research.

Bargaining

The bargaining problem for two individuals has been stated in the following way by Rubinstein (1982):

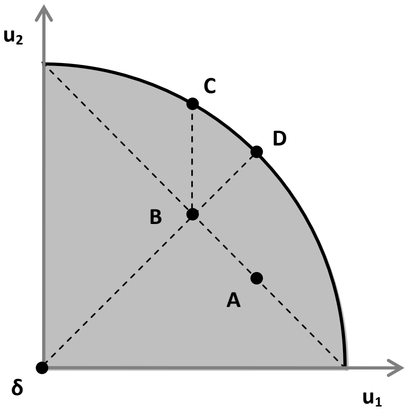
> Two individuals have before them several possible contractual agreements. Both have interests in reaching agreement but their interests are not entirely identical. What will be the agreed contract, assuming that both parties behave rationally?

A prerequisite for bargaining is that both parties will have some interest in reaching an agreement. Their potential gain from an agreement must be more than what they can achieve in their best alternative to a negotiated agreement, otherwise there is no incentive to start bargaining. Moreover, Rubinstein (1982) states that some degree of conflicting interest is a necessary requirement for a bargaining problem.

If people's interests are identical or completely shared, there is no reason to argue for a different outcome. However, a shared interest does not necessarily lead to an efficient agreement in bargaining. Lack of (truthful) communication can mean that bargainers are unable to coordinate their strategies such that they achieve the outcome that they both prefer. This coordination aspect is emphasized in the classical work of Schelling (1960) on bargaining problems. He reasoned that the bargaining outcome will depend on the coordination of parties' beliefs and showed, through a number of small experiments, that the bargaining process would converge to an outcome that is more salient than other outcomes in a particular context. Such a focal point is an outcome that stands out from the context by virtue of its simplicity, symmetry, temporal or alphabetical order, or some other feature. Norms of fairness can be focal points in bargaining.

A bargaining problem is represented in the utility that individuals obtain from the possible outcomes. In situations where the monetary gains and losses offset each other, the utilities of these gains and losses do not necessarily offset each other, for example, if one of the individuals is loss averse. Figure 1 illustrates a typical bargaining problem where the utility of individual one is shown on the horizontal axis and the utility of individual two is shown on the vertical axis. This bargaining problem could be, for example, a situation where two individuals have created a surplus from a joint venture. Disagreement arises over the ownership of the surplus, which they intend to solve by bargaining. If they disagree, the surplus is lost and they both end up with zero utility at the point marked with a δ. They both have an incentive to reach an agreement within the grey area, which is the set of all possible agreements.

Figure 1: The bargaining problem



Note: The utilities of individuals one and two are shown on the horizontal and vertical axes, respectively. Four agreements are marked $A - D$, and the disagreement outcome is marked δ. The grey area is the set of all possible agreements. The bold line is the bargaining set.

Four of the possible agreements in the grey area in Figure 1 are marked $A-D$. The bold line on the north-eastern border of the grey area is the set of all Pareto-efficient agreements. An agreement is Pareto-efficient if no other agreement exists that is strictly preferred by one player, and not less preferred by the other player. The Pareto-efficient frontier connects all efficient agreements and is called the *bargaining set.* Agreement C and D are both Pareto-efficient, and no reallocation is possible without individual one obtaining less utility. Agreements A and B represent conflicts of interest, where one individual's utility loss is offset by the other individual's utility gain. Agreement C is a Pareto-improvement to B, because individual two obtains more utility and individual one retains the same utility. Both individuals have a shared interest in moving from agreement B to agreement D because both increase their utility by doing so. Agreements B and D are egalitarian solutions where both individuals obtain the same utility.

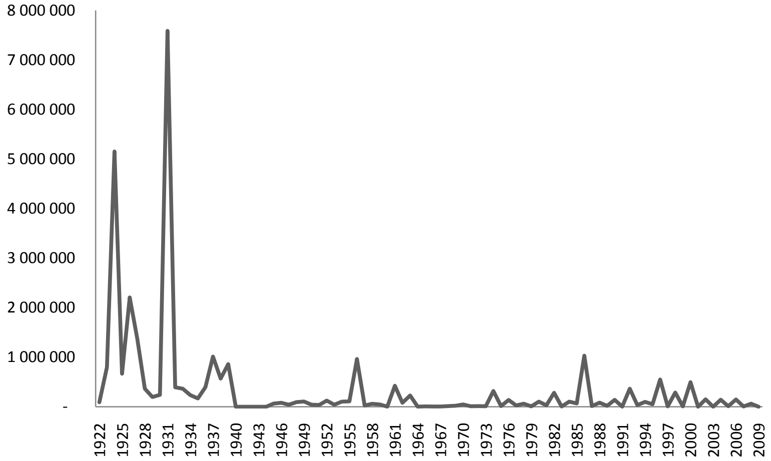
In general, the bargaining set consists of many possible outcomes. The challenge for descriptive theory is to find the most likely outcome of bargaining. Nash (1950) developed a theory that predicts a unique solution to the bargaining problem. He characterized the outcome by four axioms that he believed were reasonable premises to which a neutral third party would agree. The Nash bargaining solution applies to cooperative games where a binding agreement can be enforced, and it is independent of the bargaining process. For a bargaining situation where the parties are in symmetric positions, the Nash bargaining solution gives an equal sharing of the surplus.

Since the 1950s, a number of non-cooperative bargaining games have been developed (Harsanyi, 1977; Crawford, 1982; Rubinstein, 1982). In non-cooperative games, individuals play according to a specified bargaining protocol, that is, a set of rules that govern the bargaining process. Each individual who is playing a non-cooperative game is assumed to follow a strategy that lays out a course of action for each possible state of the game. In a bargaining setting, a strategy determines a player's offer, given the history of the game. Moreover, the strategy determines how the player responds to the offers received by other players in the game. The most well-known non-cooperative bargaining game is the alternating offer game. Under the alternating offer bargaining protocol, one of the players starts by making an opening offer. The other player can accept this offer or make a new offer to the first player. For every round of offers, the endowment shrinks by a fixed factor. The game ends when one of the players has accepted an offer. Rubinstein (1982) shows that the alternating offers game has a unique solution, which, for identical discount rates, approaches the Nash bargaining solution when the time interval gets smaller. The strength of the alternating offer bargaining protocol is its similarity to many observed negotiation processes, which, from an opening offer, typically evolve into stages where parties argue their cases and make concessions to each other until an agreement is reached.

The weakness of many game theoretic models of bargaining is the high level of abstraction that means they necessarily lack many aspects of real negotiations. In non-cooperative game theory, many details of communication are left out, for example, the ability to persuade the other party, the use of bluffing, and the misrepresentation of interests. There are some disagreements about the extent to which communications in negotiations are characterized by being manipulative or by being primarily full, open, and truthful (Raiffa, 2002). Unstructured bargaining experiments have shown that communication and social factors can easily influence the bargaining outcome. Open communication often leads to discussions about the fairness of different outcomes (Roth, 1995). Some of the references to fairness can be rationalizations of self-interest or simply cheap talk, but a concept of fairness cannot be manipulated if it is never invoked for other than manipulative purposes. For fairness arguments to play a role in communication, individuals must act upon fairness preferences at some stage (Elster, 1989). The analysis in the following essays integrates fairness into the individual's preferences, reflecting the view that fairness is more than cheap talk.

Bargaining models predict a high degree of efficiency in bargaining. The Coase Theorem states that, under the assumption of no transaction costs and complete information, the outcome of bargaining is Pareto-efficient (Coase, 1960). Data show, however, that bargaining is often inefficient, involving costly negotiations and ending in Pareto-inefficient solutions (Camerer, 2003). Figure 2 shows the number of work days lost in wage conflicts in Norway since 1922. The figure illustrates the high costs of wage negotiations in the Norwegian economy during the period before the Second World War, and the decline in costs over recent decades. Both the development of centralized labour and employer organizations, and the introduction and development of arbitration are important in wage setting in the Norwegian economy (Holden, 1998). In arbitration, a third party, often selected by the parties, has the role of a judge, and his or her award is binding for the parties.

Figure 2: Lost working days in Norway due to work conflicts



Source: Statistics Norway (Table 03629: Arbeidskonfliktar)

Arbitration is a mechanism for improving negotiation efficiency that has been extensively studied, but there is no agreement on its effects on bargaining efficiency in the literature. Since the 1960s, research has been concerned with the 'chilling effect' of arbitration on bargaining (Stevens, 1966). The argument is that if a third party compromise the final offers of the bargainers, then they are better off holding on to an extreme position rather than making concessions. Consequently, in negotiations involving the fall-back option of a third party mechanism, the conflict level increases and resources are wasted in costly negotiations and in the use of third parties. This ‘chilling effect’ has been found in a number of experiments (Ashenfelter, Currie, Farber, and Spiegel, 1992; Bolton and Katok, 1998; Charness, 2000). The central assumption involves the behaviour of the third party. Empirical studies of real arbitration awards provide mixed evidence regarding whether third parties compromise on final offers or use fairness principles (Bazerman, 1985; Bloom, 1986). In situations where fairness principles are important, a third party mechanism may facilitate the efficiency of negotiations because players hold correct beliefs about which fairness principle a third party may follow. The first essay demonstrates that negotiation under the fall-back option of a third party decision improves bargaining efficiency.

Behavioural models

The development and application of rational choice theory has been a major achievement in economics. In rational choice theory, individuals are assumed to act upon preferences that fulfil requirements about consistency of choice. More controversially, standard economic theory also assumes that preferences are only over the individual’s own material gain or loss from a transaction. In addition, rational choice theory requires assumptions about beliefs that people have about the choice situation that they are facing, and about how other people may act. The formation of beliefs is based on an individual’s current information and the seeking of relevant new information. A rational economic individual chooses, given his or her beliefs, the alternative that maximizes his or her preferences, as represented in a utility function, subject to resource constraints such as money and time.

Since the 1970s, there has been a growing number of researchers who have questioned that the assumptions of rational choice represent actual behaviour (Simon, 1983). An example of an anomaly regarding rational choice that is mentioned in the first essay of this thesis is that people tend to have a self-serving bias. In their search for information, people tend to seek information that favours their preferences. More generally, if people experience a conflict between preferences and beliefs, they will tend to adjust either their beliefs or their preferences. Cognitive dissonance is a theory that says that the adjustment should happen where the cognitive resistance is less, e.g., if your beliefs are empirically justified, the easiest way is to adjust your preferences and degrade that option (Elster, 1983).

Rational choice allows for representation of a broad set of preferences such as tastes, emotions, and norms. Research over the last few decades has extended the standard model to include other types of motivation than pure material self-interest. Three extended models that are widely discussed involve altruism, inequity aversion, and reciprocity (Becker, 1974; Bolton and Ockenfels, 2000; Rabin, 1993). An altruistic motivated individual is willing to include other individuals’ utility in his own utility function. An individual with an inequity aversion obtains disutility from outcomes that deviate from some defined principle of distributive justice, for example, equality. Reciprocity is based on the assumption that people will repay kind acts from other people and punish unkind acts from other people. Reciprocity is not related to the outcome, but directly to the other person's behaviour.

The second essay in this thesis is a contribution to the research that focuses on fairness motivations to study the bargaining problem. Here, bargaining is analysed in the situation where individuals have preferences regarding their own material outcomes and they also care about the distributive outcome of the bargaining. It is assumed that individuals obtain disutility from deviating from their preferred fairness principle. The model developed in the essay allows for individuals to adhere to different fairness principles, which is documented empirically to be important in distributive situations (Frohlich and Oppenheimer, 1992; Konow, 2003; Cappelen, Hole, Sørensen, and Tungodden, 2007). A critique of these models is that arbitrarily choosing principles can fix the model such that it can explain one phenomenon, but only with a loss of generality. The choice of which principles to include in a model is important. One source of principles is the normative political philosophy tradition; see, e.g., Rawls (1971) and Nozick (1974). The selection of fairness principles in positive economic models should be based on principles that are shown to be empirically important to people. There is a growing literature that empirically examines moral motivations, with links to the philosophical tradition (Cappelen et al., 2007; Cappelen, Sørensen, and Tungodden, 2010; Almås, Cappelen, Sørensen, and Tungodden, 2010).

The development of models in behavioural game theory that include fairness has, by a few exceptions, been restricted to players who agree on the same fairness principle (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). In the second essay, I develop a model that captures the potential conflict between two individuals who follow different fairness principles in bargaining. Fairness motivation can influence both the possibility of reaching an agreement in bargaining and it can influence the properties of the agreement that is reached. The first part of the second essay studies how fairness motivation influences the possibility of reaching an agreement. Proposition 2 formalizes the intuition of Elster (1989), that bargaining between two individuals who strongly believe in different fairness principles ends in conflict. This result shows the importance of considering a plurality of fairness principles to understand many bargaining problems. In contrast, Proposition 3 formalizes that if two bargainers follow the same fairness principle, it is always possible to reach an agreement.

Experiments

In the last three decades, we have seen an increase in the use of experiments in economic research. The use of experiments has been a catalyst for much of the development in economic models discussed in the previous section, but it is also a source of knowledge about regularities in economic behaviour that only weakly relate to specific economic models (Bardsley, Cubitt, Loomes, Moffat, Starmer, and Sugden, 2010). The major advantage of experiments is that control of the environment allows identification of the relationship between behaviour and the environment. The following essays use experiments both to test economic theory and to investigate behavioural regularities.

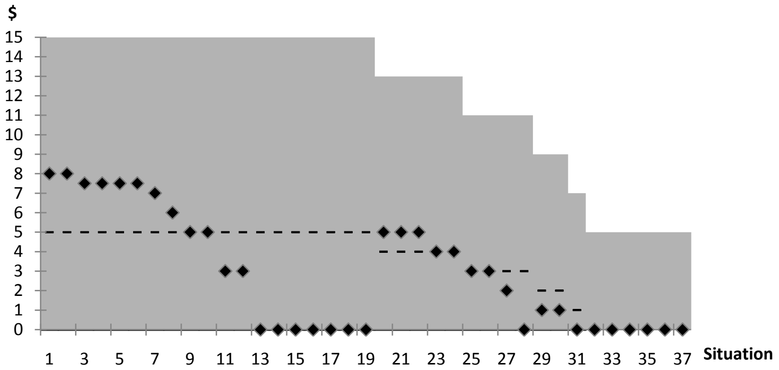
Economic experiments are normally conducted in a classroom where participants are faced with game situations in which their choices have monetary consequences. The use of monetary payoffs affects the outcomes in laboratory experiments, primarily by increasing the effort from participants, which reduces the variance of decision errors (Smith and Walker, 1993). Participants are seated in separate cubicles where all communication is restricted throughout the game. The game situation is normally described to the participant in a user-friendly neutral language to avoid cues to specific responses. These measures are undertaken to control the setting to avoid social factors influencing the results. In particular, care is taken to avoid effects from interaction between the experimenter and the participants. Characteristics of the experimenter such as gender, age, and personality have been found to influence the results of experiments. There is also a danger that 'the hypothesis held by experimenters can lead them unintentionally to alter their behaviour toward their subjects in such a way as to increase the likelihood that subjects will respond so as to confirm the investigator's hypothesis or expectations' (Rosenthal, 2002). To reduce this 'experimenter effect', a double blind procedure is used such that neither the researcher nor the other participants know a participant's choices in the experiment. Experimental procedures often allow for trial-and-error learning before or during the experiment in order to make sure that the situations are clearly understood. It is argued that people faced with unfamiliar tasks make impulsive choices and that deliberation and learning are necessary for people to discover their underlying preferences (Plott, 1996). These conventions for how economic experiments should be conducted ensure that experiments are replicable and valid. Internal validity in experimental work is assured by running all sessions identically in all respects, and drawing participants randomly from the same population into sessions. Then, researchers can make one change in the experiment design and compare the results of the two sessions. Such differences are called treatment effects and allow for causal inference.

The third essay studies the role of social preferences among a group of prisoners and a benchmark group by conducting a trust game experiment (Berg, Dickhaut, and McCabe, 1995). Trust and punishment are key norms that reduce the cost of transactions in the economy (Fukuyama, 1995). Many transactions are performed over a period of time, so that one party must voluntarily place resources at the disposal of another party (the trustee) before receiving a return and, therefore, he is exposed to the risk of not receiving the expected return

(Coleman, 1990). The trustee is trustworthy if he delivers on the expected part of the exchange. The decision to place trust in another party depends on the individual's subjective beliefs about the chances that the other party will break the trust. Misplaced trust results in a loss, but the failure to place trust in a trustworthy party may also have a high cost in terms of gains forgone.

The following data are from a classroom experiment of the trust game, and illustrate how an economic experiment can be used to study trust. Students who attended courses in experimental economics at the Norwegian School of Economics and Business Administration over the years 2006–2008 participated in three sessions. The basic design of the trust game experiment was that participants were randomly matched into pairs. There were 37 students in the role of senders and another 37 students in the role of receivers. Senders got an endowment of $5 from the experimenter and could send zero or increments of one $ to the receivers. The amount that was sent to the receivers were tripled and they had to decide whether to return all or some whole $ portion of the tripled amount to the sender. The maximum earnings in this game were $15, and the minimum earnings occurred when the sender kept his $5.

Figure 3: Trust game experiment results



Note: The broken line represents the sent amount, the grey bar is the total amount available to the receiver, and the black diamond represents the amount returned by the receiver for each of the 37 game situations.

The highest monetary rewards can be achieved by sending all the money to the other player. The problem is that the sender cannot be sure that the receiver will return anything and, therefore, he will be better off keeping all the money for himself, which is the prediction of standard economic theory. Figure 3 shows the results from the student experiment. On average, the participants were trusting, sending 70% of the endowment. However, there were significant behavioural variations around this average, with more than half of the students sending the whole endowment, and six students keeping everything. The average returned amount was about 80% of the sent amount, but here also there were variations, with the returned amount ranging from an equal sharing of the total earnings to zero. We can see that some observations cluster around a returned amount that equals the sent amount. In the trust game, on average, the sender earns less than his or her endowment, whereas the receiver keeps the surplus generated from the transaction. This pattern is quite typical for trust game results. The trust game experiment (and other frequently used games) have been shown to produce results that are robust for higher endowments, for repetition, and for non-student populations (Camerer, 2003).

The traditional interpretation of this experiment is that the sender amount measures the sender's trust and that the returned amount measures the trustworthiness of the receiver. The trust game is considered an adequate way to operationalize the phenomena of trust. However, it has also been argued that the trust game does not measure trust, because there is no promise from the receiver involved, and that, instead, the trust game measures willingness to undertake risky investments (Bohnet and Zeckhauser, 2004). The standard trust game can uncover empirical regularities between different subject pools, which suggest that there are some real differences in behaviour between the subject pools. However, the standard trust game does not reveal the motivations of the participants. People may show trust in this game because they are self-interested and expect the receiver to return money, because they have preferences for an increase of total income, or because they are altruistic. To identify motives in a trust game, it is necessary to change the experiment design or to combine the results with other observations regarding the same individuals (Cox, 2004; Glaeser, Laibson, Scheinkman, and Soutter, 2000).

Recent economic experiments with games, such as the dictator game and the trust game, have documented that social preferences are important in explaining behaviour in situations where decisions have consequences for others, and have also indicated that there are considerable differences in social preferences both within and across groups (Henrich, Boyd, Bowles, Camerer, Fehr, and Gintis, 2004). In the third essay, we report results from an experiment in which we compare the behaviour of a group of prisoners with a benchmark group recruited from a representative sample of the Norwegian population. The experiment consists of a dictator game and two versions of the trust game: a standard trust game and a trust game with punishment. Heterogeneity in the importance attached to pro-social preferences could potentially be important in explaining criminal behaviour because, typically, crime has negative consequences for others. More specifically, if people take into account how their actions affect others before they decide whether or not to commit a crime, then the likelihood of a person committing a crime would be decreasing in the importance he attaches to pro-social preferences. Consequently we would expect criminals on average to be less motivated by pro-social preferences than non-criminals. We find, however, that the prisoners are highly motivated by pro-social preferences, both in strategic and non-strategic situations, and that there is a striking similarity in the importance that the prisoners and the benchmark group attach to pro-social preferences.

Group identity has been shown to be important for social preferences, and both in-group favouritism and out-group discrimination are important phenomena in some contexts (Akerlof and Kranton, 2010). As prisoners could possibly identify with the other prisoners in the experiment, there could be an in-group effect on their social preferences. Therefore, in addition to single-group sessions, where participants only interacted with participants from their own group, we included mixed-group sessions, where participants interacted with participants from both groups. This allowed us to study how prisoners behaved when they interacted with non-prisoners. In addition, the mixed sessions allowed us to study whether the benchmark group was prejudiced against prisoners. However, we found little evidence of in-group favouritism or out-group discrimination. This result suggests that prisoners do not identify strongly with the general prison population.

In the second essay, a bargaining experiment illustrates the influence of fairness on the bargaining outcome, which is used to develop a model of bargaining behaviour by including a plurality of fairness principles. There is a long tradition in experimental economics of studying bargaining behaviour, fairness, and different theoretical solution concepts for bargaining (Fouraker and Siegel, 1963; Ochs and Roth, 1989; Weg, Rapoport, and Felsenthal, 1990; Binmore, Swierzbinski, Hsu, and Proulx, 1993). The experimental procedure used in the second essay is a direct implementation of the alternating bargaining protocol that is used in theoretical bargaining models. Because the experiment implements the assumptions of the theoretical model, we should expect that the experiment is clearly within the domain of the theory, and that if the predictions of the model fail, this is due to the assumptions of the model. Specifically, if bargaining theory claims to be general, its predictions should also hold for contexts with production, as in the second essay. A problem with this type of argument is that all implementation of theory requires some sort of auxiliary hypothesis. For example, economic theories are based on assumptions about utility functions, whereas the implementation in experiments uses monetary rewards, which requires some auxiliary hypothesis about the mapping from monetary rewards to utilities. It is, therefore, not clear in this example if the results from the experiment mean that the bargaining solution concept fails, or that the mapping of monetary awards into utilities fails. Viewed in isolation, this is a problem for all applied economic theory.

Bardsley et al. (2010) suggest that experiments on social preferences should be interpreted within a broader research programme called the ‘preference refinement program’, where the predictive success of particular solution concepts is part of the hard core and hypotheses about preferences are treated as open to adjustments. This contrasts with the ‘applied game theory program’ that takes hypotheses about preferences as part of its hard core and treats claims about the predictive success of particular solution concepts as open to adjustments.

Bardsley et al. (2010) argue that the individual experimental results should be evaluated in relation to the progress of the broader research program.

In the first essay, an experiment is used in a slightly different way to investigate the effect of a possible third party decision on the efficiency of bargaining. The experimental literature on arbitration has used an approach where the experimental designs mimic the various rules that a third party could apply in a real arbitration, such as conventional arbitration, final offer arbitration, tri-offer arbitration, etc. (Ashenfelter et al., 1992). The effects of these rules can be studied within the experiment, and the results can be used to improve the design of real arbitration. In some respects, this approach to experiments can be more informative than theoretical models, and it has been used successfully in the development of auctions (Smith, 2008).

A much-discussed issue within economics is to what extent the results from laboratory experiments can be generalized to other contexts, especially to less-controlled interactions. The extent to which experimental results allow for conclusions to be formed about behaviour outside the experiment is called external validity. A typical claim against external validity is that the artificiality of the laboratory environment creates behaviour that is not seen in the field or an uncontrolled environment (Levitt and List, 2007). To some extent, the laboratory experiment may introduce artificial situations that would not generalize to other environments, but the same problem exists for any empirical method. The selection of an appropriate method in economics for testing hypotheses and explaining behaviour should be a pragmatic choice based on the standards of the research community.

References

Akerlof, George A. and Rachel E. Kranton (2010). *Identity Economics*, Princeton University Press.

Allingham, Michael G. and Agnar Sandmo (1972). "Income tax evasion: A theoretical analysis", *Journal of Public Economics*, 1: 323–338.

Almås, Ingvild, Alexander W. Cappelen, Erik Ø. Sørensen, and Bertil Tungodden (2010). "Fairness and the development of inequality acceptance", *Science*, 328 (5982): 1176–1178.

Andvig, Jens Chr. and Karl Ove Moene (1990). "How corruption may corrupt", *Journal of Economic Behaviour and Organization*, 13: 63–76.

Ashenfelter, Orley, Janet Currie, Henry S. Farber, and Matthew Spiegel (1992). "An experimental comparison of dispute rates in alternative arbitration systems", *Econometrica*, 60(6): 1407–1433.

Bardsley, Nicholas, Robin Cubitt, Graham Loomes, Peter Moffat, Chris Starmer, and Robert Sugden (2010). *Experimental Economics : Rethinking the Rules*, Princeton University Press.

Bazerman, Max H. (1985). "Norms of distributive justice in interest arbitration", *Industrial and Labor Relations Review*, 38(4): 558–570.

Becker, Gary S. (1968). "Crime and punishment: An economic approach", *Journal of Political Economy*, 76(2): 169–217.

Becker, Gary S. (1974). "A theory of social interaction", *Journal of Political Economy*, 6: 1063–1091.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995). "Trust, reciprocity, and social history", *Games and Economic Behavior*, 10(10): 122–142.

Binmore, Ken, Joe Swierzbinski, Steven Hsu, and Chris Proulx (1993). "Focal points and bargaining", *International Journal of Game Theory*, 22: 381–409.

Bloom, David E. (1986). "Empirical models of arbitration behavior under conventional arbitration", *The Review of Economics and Statistics*, 68: 578–585.

Bohnet, Iris and Richard Zeckhauser (2004). "Trust, risk and betrayal", *Journal of Economic Behaviour and Organization*, 55: 467–484.

Bolton, Gary E. and Elena Katok (1998). "Reinterpreting arbitration's narcotic effect: An experimental study of learning in repeated bargaining", *Games and Economic Behavior*, 25: 1–33.

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

Camerer, Colin (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press.

Cappelen, Alexander W., Astrid D. Hole, Erik Ø. Sørensen, and Bertil Tungodden (2007). "The pluralism of fairness ideals: An experimental approach", *American Economic Review*, 97(3): 818–827.

Cappelen, Alexander W., Erik Ø. Sørensen, and Bertil Tungodden (2010). "Responsibility for what? fairness and individual responsibility", *European Economic Review*, 54(3): 429–441.

Charness, Gary (2000). "Bargaining efficiency and screening: an experimental investigation", *Journal of Economic Behaviour and Organization*, 42: 285–304.

Coase, Ronald H. (1960). "The problem of social cost", *The Journal of Law and Economics*, 3: 1–44.

Coleman, James S. (1990). *Foundations of Social Theory*, The Belknap Press of Harvard University Press.

Cox, James C. (2004). "How to identify trust and reciprocity", *Games and Economic Behavior*, 46: 260–281.

Crawford, Vincent P. (1982). "A theory of disagreement in bargaining", *Econometrica*, 50: 607–38.

Elster, Jon (1983). *Sour Grapes – Studies in the Subversion of Rationality*, Cambridge University Press.

Elster, Jon (1989). *The Cement of Society*, Cambridge University Press.

Fehr, Ernst and Klaus M. Schmidt (1999). "A theory of fairness, competition and cooperation", *The Quarterly Journal of Economics*, 114(3): 917–868.

Fouraker, Lawrence E. and Sidney Siegel (1963). *Bargaining Behavior*, McGraw-Hill.

Frohlich, Norman and Joe A. Oppenheimer (1992). *Choosing Justice. An Experimental Approach to Ethical Theory*, University of California Press.

Fukuyama, Francis (1995). *Trust*, Free Press.

Glaeser, Edward L., David I. Laibson, José A. Scheinkman, and Christine L. Soutter (2000). "Measuring trust", *The Quarterly Journal of Economics*, 115(3): 817–868.

Harsanyi, John C. (1977). *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis (2004). *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford University Press.

Holden, Steinar (1998). "Inntektspolitikken – hvordan virker den og hva kan oppnaas?", Working paper Norwegian Ministry of Finance, no. 29.

Johansen, Leif (1979). "The bargaining society and the inefficiency of bargaining", *Kyklos*, 32: 497–522.

Konow, James (2003). "Which is the fairest one of all?", *Journal of Economic Literature*, XLI: 188–1239.

Levitt, Steven D. and John A. List (2007). “What do laboratory experiments measuring social preferences reveal about the real world”, *Journal of Economic Perspectives*, 21: 153–174.

Nash, John F. (1950). “The bargaining problem”, *Econometrica*, 18: 155–62.

Nozick, Robert (1974). *Anarchy, State, and Utopia*, Basic Books.

Ochs, Jack and Alvin E. Roth (1989). “An experimental study of sequential bargaining”, *American Economic Review*, 79(3): 355–384.

Plott, Charles R. (1996). “Rational individual behaviour in markets and social choice processes: the discovered preference hypothesis”, in “The Rational Foundations of Economic Behavior”, Macmillan, (pp. 225–250).

Rabin, Matthew (1993). “Incorporating fairness into game theory and economics”, *American Economic Review*, LXXXIII(5): 1281–1302.

Raiffa, Howard (2002). *Negotiation Analysis : The Science and Art of Collaborative Decision Making*, The Belknap Press of Harvard University Press.

Rawls, John (1971). *A Theory of Justice*, Harvard University Press.

Rosenthal, Robert (2002). “Experimenter and clinician effects in scientific inquiry and clinical practice”, *Prevention & Treatment*, 5: Article 38.

Roth, Alvin E. (1995). “Bargaining experiments”, in “The Handbook of Experimental Economics”, Princeton University Press, (pp. 253–348).

Rubinstein, Ariel (1982). “Perfect equilibrium in a bargaining model”, *Econometrica*, 50: 97–109.

Schelling, Thomas C. (1960). *The Strategy of Conflict*, Harvard University.

Simon, Herbert A. (1983). *Reason in Human Affairs*, Stanford University Press.

Smith, Vernon L. (2008). *Rationality in Economics : Constructivist and Ecological Forms*, Cambridge University Press.

Smith, Vernon L. and James M. Walker (1993). “Monetary rewards and decision cost in experimental economics”, *Economic Inquiry*, 31: 237–244.

Stevens, Carl M. (1966). “Is compulsory arbitration compatible with bargaining”, *Industrial Relations*, 5: 38–52.

Weg, Eythan, Amnon Rapoport, and Dan S. Felsenthal (1990). “Two-person bargaining behavior in fixed discounting factors game with infinite horizon”, *Games and Economic Behavior*, 2: 76–95.

Negotiation under possible third party settlement[*]

Sigbjørn Birkeland d.y.[†]
Department of Economics
Norwegian School of Economics and Business Administration

15th November 2010

Abstract

The effect of possible third party settlement on negotiation behaviour is studied in an economic bargaining experiment. The bargaining phase is preceded by a production phase that allows for different fairness principles to guide the division of the total production value. The experimental results show that a possible third party settlement lowers the dispute costs by reducing the number of rounds of alternating offers. In the presence of a third party, negotiators make first offers that are more strongly related to their production, which reduces the number of rounds of bargaining. The production phase has an effect on the distributional property of the settlements. In negotiations where third party settlement is an option, the negotiation outcome shifts towards a more unequal outcome, more in line with each person's contribution.

Keywords: Arbitration; Bargaining efficiency; Experiment
JEL classification: C78; D63; J52

[*] I would very much like to thank Bertil Tungodden, James Konow, Alexander Cappelen, and Erik Ø. Sørensen for valuable comments and helpful discussions, and Øystein Dvergsdal, Ole-Aleksander Greve, and Maria Tsujimoto Frengstad for research assistance. Thanks also to participants at a seminar at the University of Bergen, the conference New Directions in Welfare at St. Catherine's College, Oxford University, and the 4th Nordic Conference on Behavioral and Experimental Economics at the University of Oslo. The experiment costs were financed by the Research Council of Norway, research grant 185831, and the Department of Economics at the Norwegian School of Economics and Business Administration, Bergen.

[†] Address: Helleveien 30, 5045 Bergen, Norway. E-mail `sigbjorn.birkeland@nhh.no`.

1 Introduction

Many decisions are reached in negotiations under the fall back of a third party settlement. Civil disputes can be brought to court, disputes arising under commercial contracts may be solved by arbitration proceedings, and conflicts between branch managers can be decided by a senior manager. The main question addressed in this paper is to what extent the possibility of submitting a case to an impartial third party for settlement influences the bargaining efficiency and the distributional properties of the settlements. The effects are studied in a laboratory experiment with business school students. The negotiations studied in this paper are such that two parties must agree upon the division of a sum of money created through individually produced output. The bargaining process is costly and may go on until no money is left on the table. I compare negotiations with and without the option to unilaterally submit the case to an impartial third party. There are three possible outcomes in the game: an agreement about a division of the money, a final third party settlement, or a perpetual disagreement. The game studied is based on the alternating offer bargaining protocol extended to include the outside option of using a third party to make a final decision.

Experimental studies of sequential bargaining show that concerns for fairness influence bargaining behaviour (Ochs and Roth, 1989; Weg, Rapoport, and Felsenthal, 1990; Bruyn and Bolton, 2008). In these experiments the players are asked to negotiate the division of a fixed amount of money over a few rounds of offers. The experimenter induces differences in the individual discount rates which, according to the standard model, should give an unequal division of the money. The experimental results show, however, that players tend to favour an equal distribution of money. Even in experimental situations where one of the negotiators has all the bargaining power (dictator game), a third of the participants typically divide equally (Camerer, 2003). The strong tendency for an equal split in these bargaining contexts may be driven by the widely accepted fairness principle of an equal split when no entitlement to the money exists. The experiment reported in this paper is designed to create different entitlements to the endowment through a real effort production phase before negotiations take place. The experimental results show that a player with a higher production gets on average a larger share of the money. In the presence of a third party, the distributional property of the settlements shifts towards a more unequal outcome that is more in line with each person's contribution.

Experimental studies of arbitration typically find that dispute rates more than double when conventional arbitration is introduced into negotiations (Ashenfelter, Currie, Farber, and Spiegel, 1992; Bolton and Katok, 1998; Charness, 2000; Dickinson, 2004).[1] In the experiment reported here the introduction of an option to let a third party decide significantly lowers the dispute costs by reducing the

[1] With the exception of Charness (2000), these other studies do not apply the alternating offer protocol and the negotiation process is not costly, but money is lost in the case of disagreement.

number of rounds of alternating offers. The reduction in the number of rounds of bargaining can be explained by negotiators making first offers that are more strongly related to their production when in the presence of a third party. The introduction of a third party therefore influences both the efficiency and the distributional properties of negotiations.

More details of the experimental design are presented in Section 2. Section 3 contains a theoretical analysis based on standard bargaining models. The experimental results are discussed in Section 4, and the relation to the experimental bargaining literature is provided in Section 5. Some concluding remarks are given in Section 6.

2 Experimental design

The experiment contains four phases: a production phase, a dictator phase, a negotiation phase, and a question phase. Participants are provided with the basic design of all four phases at the beginning of the experiment (complete instructions are provided in Appendix A). There are two experimental treatments, and the participants are randomly selected into one of the two treatments. Those that act as third parties are also randomly selected among the participants. Third parties do not participate in the production or dictator phase in order to not bias their view. Instead, they spend their time answering questions on four hypothetical cases that are similar in structure and information to the real cases that they meet later in the experiment.

In the dictator and negotiation phases, the participants are randomly matched in pairs and each person is involved in four situations in the dictator game and four negotiations. Pairs are rematched between each situation. Participants in each session are randomly seated in separate cubicles; all interaction between participants is anonymous and made through a web interface developed for the experiment (selected screenshots are provided in Appendix C).[2]

Payment in the experiment is determined for each participant by a random draw from the four situations in the dictator phase or the four situations in the negotiation phase that the participant has been involved in. The participant is paid according to the result he or she achieved in the situation that is drawn. If there is a third party involvement, the participant is paid according to the third party decision net of third party costs. A third party is paid a fixed amount of compensation by the experimenter, independent of his or her choices in the experiment or whether the service is used at all. In many third party institutions

'Dispute rates' refers to the fraction of bargains where no agreements have been made after a fixed time period.

[2] The experiment is programmed in Python, and uses a MySQL database and an Apache web server application. The experiment is run on laptop computers that communicate over a wireless local area network.

such as commercial arbitration, a third party is normally paid by the parties for the time used to settle the dispute. The choice of a fixed sum communicated to the third parties upfront was made in order to make incentives clear and unbiased with respect to the uncertain demand for their service in the experiment.

At the end of the experiment, each participant is asked to complete a form using a code given on the screen and the payment attached to that code is transferred to the participant's bank account. Matching of the receivables and the bank account details is done by a person outside the research group who has no other information about the experiment.

In the *production* phase of the experiment all negotiators produce individually an output by copying text for 10 minutes on the computer. The production phase has been designed to create individual entitlements to the money. Individual production is rounded off to the nearest 50 correct words typed. Individual i's production value y_i is equal to $e_i p_i$, where the number of words typed is e_i, and the price p_i is either NOK 0.75 or 1.50 per correct word.[3] The prices are randomly distributed to players by the experimenter after typing has ended.[4] The total production value to be divided in a negotiation is equal to $Y = y_1 + y_2$.

There are at least three salient fairness principles, m^n where $n = E, L, P$, which can guide the individual in dividing the total production value. The first principle is strict *equality* which is simply an equal split of the joint production value, $m^E = Y/2$. The second principle is a *laissez-faire* principle which gives each individual what he earns in the production $m^L = e_i p_i$, and the third principle is a *proportionality* principle which allocates the joint production according to the relative production of words such that individual i gets $m^P = (e_i/(e_1 + e_2))Y$. A principle of proportionality under which the input–output ratio is equal between people is often called the equity principle or the accountability principle (Konow, 1996). It is a widely held principle, especially in contexts of production (Konow, 2000).

In the *dictator* phase of the experiment, participants are randomly matched in pairs and one participant is chosen to act as a dictator who decides on the division of the production value, Y, between the two. The participant acting as a dictator is given full information about both participants' production of words and the randomly assigned prices. Each participant is involved in a total of four dictator situations, two as a dictator and two as a passive receiver, all randomly matched pairs. The dictator game represents a situation comparable to negotiations where one of the players has all the bargaining power and there is no strategic element to the distributive choice. Information from the dictator situations is used to explain behaviour in the negotiations.

[3] At the time of the experiment USD/NOK = 6.9

[4] Before the production phase all participants are told that they will earn money according to the number of correct words that they type, but that the payoff from the experiment will depend on the subsequent phases. To avoid incentive effects, prices are assigned after the production.

In the *negotiation* phase of the experiment, participants are again randomly matched in pairs and they are instructed to bargain over the division of the production value, Y. The bargaining protocol is an alternating offer bargaining with infinite horizon.[5] Both participants are induced with an equal discount factor, $\delta = 0.96$, such that the value of the production shrinks by the same amount for both negotiators. Money at the negotiation table in round t is equal to $\delta^{t-1}Y$. One of the participants in a pair is randomly assigned as the first mover and proposes an opening offer in the first round, $t = 1$. An offer from individual i is an amount of money x_i to himself and $Y - x_i$, to the other party. In treatment I, called the *bargaining* treatment, the second mover responds to an initial offer either by accepting it and the negotiation is closed without cost, or by making a counteroffer in the second round ($t = 2$) where the production value is reduced by 4%. The negotiation is closed when one party accepts an offer.[6] The pairs of participants are rematched between each negotiation and all players take part in four negotiations. Participants are given full information about the other participants' production of words and the prices assigned to each in the bargaining pair. Every offer that is made during the negotiation is recorded in a table on the screen. Communication between parties is restricted to this minimal exchange of suggested divisions of the total production value, and acceptance or rejection of the other's offer.

In treatment II, called the *third party* treatment, the bargaining protocol is changed such that there is an additional option available during the alternating offers. This is to unilaterally break off the negotiation and request a third party settlement. Because this extra third party option is only available in the second treatment and there is random assignment of participants to the two treatments, the experimental design allows us to study the causal effect of introducing a third party option. Using a third party costs each negotiator 5%. The cost is independent of who made the request for the third party settlement. There are no restrictions on the settlement imposed by the third party other than it has to be equal to the available sum of money, so that no money can be added or withdrawn. The third party called upon to make a decision is given all the relevant information about the negotiation, that is, both negotiators' production of words, the assigned prices, and the full sequence of offers made by both negotiators including who asked for the third party service. Negotiators are not given

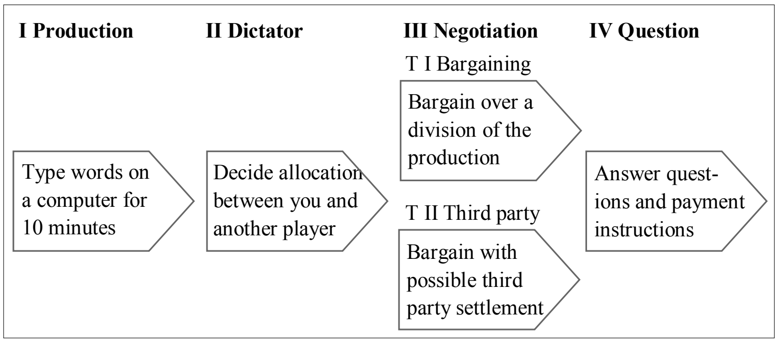
[5] In principle, the negotiations could continue until the minimum offer of NOK 1 is reached or the participants could use an excessively long time to decide in each round, never concluding the negotiation. From previous experience with experiments of this kind, we thought these events so unlikely that the participants were not informed of how such situations would be handled. One negotiation lasted for 26 rounds ending with an equal split of the remaining 0.36% of the production. It took 22 minutes to complete this negotiation.

[6] Every time a choice has been made, participants are informed about the consequences of their choice, and they are asked whether they would like to revise it before it is transmitted to the other party.

information about any decisions made by third parties during the experiment.

After all choices are made, the participants are given three questions about bargaining and fairness (questions can be found in Appendix B). Figure 1 shows the different phases and the two treatments of the experiment for the negotiators.

Figure 1: Experimental design for negotiators


I Production
II Dictator
III Negotiation
IV Question
T I Bargaining
Bargain over a division of the production
Type words on a computer for 10 minutes
Decide allocation between you and another player
T II Third party
Bargain with possible third party settlement
Answer questions and payment instructions

Beliefs about the potential outcomes of the negotiations and the third party decisions are elicited during the experiment. This allows for checking whether the outcomes are affected by mistaken beliefs about the outcome of negotiations, for example whether the use of a third party is driven by mistaken beliefs about the third party decision. Before the first mover sends the initial offer, he is asked what he believes will be the outcome of the bargaining. The first mover receives a bonus of NOK 20 if the guess is within a NOK 20 deviation of the actual agreement made.[7] In all cases where a third party settlement is requested, the participant who requests the third party settlement is asked what he thinks is the most likely outcome. Participants are paid a bonus of NOK 20 if the answer is within a NOK 20 deviation of the actual decision made by the third party.

The experiment took place at the Norwegian School of Economics and Business Administration in October 2008. Students from the first and second years of the Master of Science programme in Economics and Business Administration were invited to participate in an experiment. The invitation explained that the experiment was voluntary, that they would receive NOK 100 for participating, and that they would possibly earn more money during the experiment. A total of 110 students volunteered to participate and they were randomly assigned to one of six sessions, three sessions for each treatment. There were 28 bargaining pairs in the pure bargaining treatment and 24 bargaining pairs in the third party treatment. The 104 negotiators were paid an average of NOK 333 (USD 48.3) for an experiment that lasted on average an hour and a half. The maximum payment any student received was NOK 600. For the third party treatment, six

[7] To avoid any strategic behaviour with respect to final offers and the bonus payment, it is made explicit in the instructions that a negotiator will not receive the bonus if that particular situation is drawn for payment.

of the students were randomly selected to act as a third party. Third parties were paid a fixed compensation of NOK 300.

3 Theoretical analysis

This section discusses what the expected difference between the two treatments should be based on standard models of bargaining behaviour. The negotiation protocol used in the experiment is based on an alternating offer model with an infinite horizon that has a unique sub-game perfect equilibrium outcome (Rubinstein, 1982). In the absence of a third party decision, the model predicts that the first mover will offer $\delta/(1+\delta)$ to the other player, who should accept. In the experiment, δ is induced to be 0.96 for both players. The first mover should therefore offer 0.49 to the other participant, who should then accept. A low discount rate of 4% is chosen in order to reduce the first mover advantage. The Rubinstein model predicts an outcome of the negotiations close to an equal split with no variation. An agreement made in the first round is costless and efficient. The model is based on both players having standard preferences, which are common knowledge among the players. For small payoffs the utility function can be assumed to be linear in payoffs representing risk neutrality. The production phase does not enter into the model, which is based on a given endowment to negotiate over.

The influence of a third party on the negotiations will depend upon the rules that govern the third party mechanism and the assumptions about the third party behaviour. Here, negotiators can unilaterally submit the case for a binding third party decision in any round during the negotiations. A third party can implement any settlement of the contested amount, but he cannot add or subtract money. There are different hypotheses about how impartial third parties reach decisions. Many papers on arbitration assume that the arbitrator will compromise the final positions of the negotiators. Negotiators will in such a situation tend to make large demands and small concessions in order to offset the compromise decision of the arbitrator. Such behaviour would predict increased dispute rates in negotiations under a possible third party settlement. On the other hand, if a third party follows a fairness principle in the allocation decision, the effect on dispute rates may be different. Among the papers that study arbitrator behaviour empirically, mostly in labour disputes, Bazerman (1985) finds that arbitrators consistently apply principles in the decisions across different cases, and that there is variation among arbitrators in which principle they apply, while Bloom (1986) finds more evidence of compromising behaviour among arbitrators.

Submitting the case to a third party for settlement is an outside option. In situations where negotiators know with certainty what principle the third party will use, a rational negotiator with standard preferences should submit the case for a third party decision when the payoff from a third party settlement net of costs is greater than the payoff that would be the outcome of a negotiation. Because the Rubinstein model predicts an almost equal split, the outside option will be an empty threat for both players if it is common knowledge that the third party follows an *egalitarian* principle. This is because of the cost of using a third party. Hence, if the Rubinstein model predicts correctly the behaviour of the players, there should be no difference between the treatments; all negotiations should end in the first round with an equal split of the money.

If the third party is known to follow either a *proportionality* or a *laissez-faire* principle, the outcome will depend upon the application of these principles in the specific situation facing the negotiators in the experiment. In a few cases the production and the price are the same for both negotiators, and an application of any of the fairness principles will then lead to the same answer—an equal split. However, in most situations where there are differences in the number of words produced or the prices assigned, the application of these principles gives more money to one of the parties. This party could then use this as a credible threat to get more money out of the negotiation. The other party should recognize the credible threat and agree on a settlement that follows the principle of the third party. Hence, if both negotiators know that the third party follows a *proportionality* or *laissez-faire* principle, we should expect differences between the treatments. The distributional properties of the bargaining outcome should on average be more unequal, reflecting the fact that self-interested negotiators have a credible outside option threat.

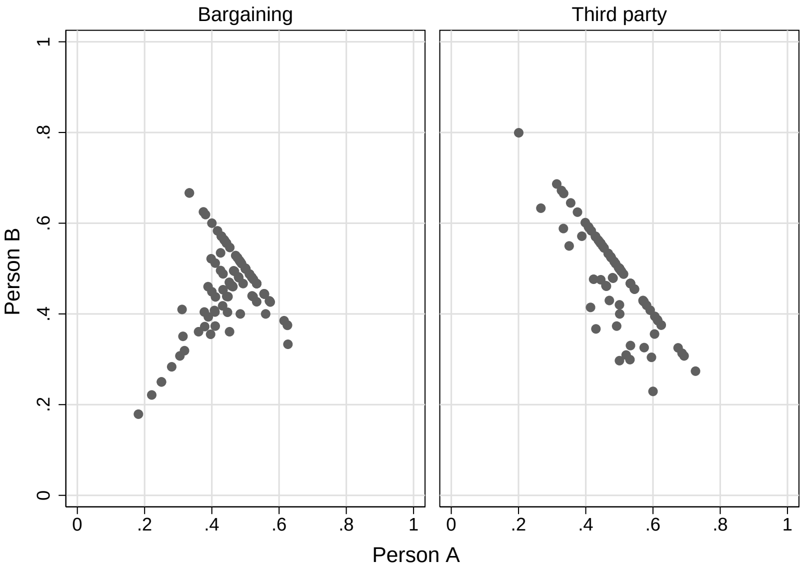
If there is uncertainty about the third party decision then the negotiators would take this into account. There is an expected gain from submitting the case to a third party if the expected outcome net of cost is higher than the outcome from a bargaining agreement. Uncertainty about third party principles should not in itself change the conclusion about the expected differences between the treatments.

Rational negotiators with perfect information should agree in the first round, independently of their preferences and the presence or absence of a third party. An agreement in the first round is Pareto efficient. A third party may influence the distribution of the negotiation outcome, but he should not influence bargaining efficiency. Third party arbitration is a costly mechanism; the threat of using it should be sufficient to influence the outcome and no actual use should therefore be observed. However, in many experiments negotiators use multiple rounds of offers and counteroffers to reach an agreement. Such inefficiencies in negotiations could arise from bounded rationality and uncertainty. The Rubinstein (1982) solution relies on rationality in the sense that the parties should be able to solve the problem using backward deduction. It is however well known that participants in experiments, for example the centipede game, fail in the use of backward deduction logic (Camerer, 2003). Uncertainty with respect to the other players' preferences or motives could also create more rounds of negotiations because negotiators use costly delays to signal to the other party information about their own reservation value (Ochs and Roth, 1989). Because participants are randomly assigned to the two treatments, bounded rationality and signalling are not expected to cause differences in efficiency between the treatments.

4 Experimental results

The experimental results show that both the bargaining efficiency and the distributional properties of the outcome are influenced by the introduction of a possible third party settlement. The 208 negotiations are summarized for the two treatments in Figure 2. Each point on the graph on the left represents an agreement from the bargaining treatment with person A's share of the total production value on the horizontal axis and person B's share on the vertical axis. Each point on the graph on the right represents a settlement from the treatment with an option to submit the case to a third party, including 15 actual third party decisions. All the points along the diagonal line from the upper left corner to lower right corner represent efficient agreements, i.e. agreements made without costs. All the points that are placed inside this efficiency frontier represent settlements where some of the production value is lost during negotiation or by the use of a third party. We can immediately observe that more settlements from the bargaining treatment are further away from the efficient frontier, indicating a difference in efficiency between the treatments.

Figure 2: Share of total production value



Almost half of all the settlements are equal splits.[8] There is a difference be-

[8] In order to accommodate rounding to the nearest NOK 5 and the small first mover ad-

tween the treatments: 56% of the settlements in the bargaining treatment are equal splits compared with 34% of the settlements in the third party treatment. There is also a significant increase in the variance of the share that person A receives in the third party treatment ($p < 0.001$, Levene's test). The most extreme unequal split is a 20–80 split from the third party treatment.[9]

4.1 Efficiency in reaching an agreement

Efficiency in reaching an agreement is measured in terms of the reduction in the production value and actual use of a third party. Dispute cost, c, in the cases where no third party is used, is equal to one minus the accumulated reduction in value from the discount factor ($c = 1 - \delta^{t-1}$, where t is the number of rounds at the close of an agreement). In the cases where a third party is used, dispute costs are also adjusted for the third party cost, α, of 10% ($c = 1 - \delta^{t-1}(1 - \alpha)$). If the second mover accepts the initial offer in the first round, there will be no dispute cost.

Table 1: Efficiency of settlements by treatment

		I: Bargaining	II: Third party
Rounds	average	3.29	1.33
	std. dev.	(4.41)	(0.75)
Dispute costs	average	0.08	0.03
	std. dev.	(0.13)	(0.05)
n		112	96

Table 1 shows that dispute costs are significantly reduced from 8% to 3% in the third party treatment ($p = 0.004$, Wilcoxon rank-sum test). The average number of rounds in the third party treatment is 1.3, which is significantly lower than 3.3 in the bargaining treatment ($p < 0.001$, Wilcoxon rank-sum test). In the bargaining treatment, about 20% of the negotiations continue for five rounds or more; see Figure 3.

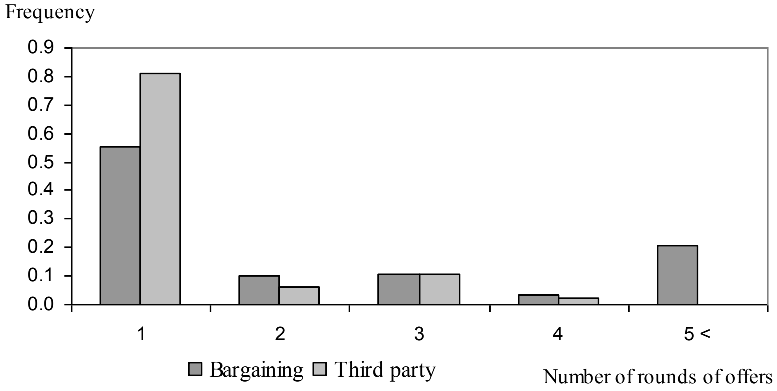
The results show that there is a significant improvement in efficiency by making available an option to submit the case to a third party. This is in contrast with previous experimental studies that find an increase in dispute rates when a third party is introduced (Ashenfelter et al., 1992; Bolton and Katok, 1998;

vantage that follows from a theoretical solution, all agreements within a 47.5–52.5 split are characterized as equal splits. Forty-six per cent of the settlements are within this bound.

[9] The average split is close to 50–50 for both treatments. Because of the random selection of pairs in the experiment, a consistent application of any of the principles discussed in Section 3 gives an average of 0.5, but with different variance. The distributional properties of the agreements are further discussed in Section 4.3.

Charness, 2000; Dickinson, 2004). These studies typically find that dispute rates more than double when conventional arbitration is introduced. (There are some differences in experimental design between these papers that can possibly explain the different dispute rates; this is discussed in Section 5.)

Figure 3: Rounds of offers in the two treatments



The first offer is important for the negotiations because it sends a signal about the preferred outcome and the aggressiveness of the strategy that a negotiator will employ. The first offers could therefore have a strong impact on the efficiency of negotiations. The number of rounds that a negotiation takes to complete in the experiment is significantly correlated with the first offers ($p < 0.001$, Spearman rank-order correlation). The main explanation of the reduction in dispute costs is the sharp decrease in rejections of first offers when introducing a third party (45% versus 19%).

There is a notable difference in the average share that person A offers in the first round in the two treatments. The average share that person A offers in the first round is 8% higher for the third party treatment compared with the bargaining treatment, which is a statistically significant difference ($p = 0.10$, Wilcoxon rank-sum test). The higher offers in the first round, when there is an option to submit the case to a third party, indicate that there is a difference in the negotiation strategies employed in this treatment.

In Table 2 the first offer from person A is explained by his relative production in the pair, and his relative price in the pair. The coefficient for production would be equal to one if the outcome was proportional to the production of the negotiators, that is, if the proportionality principle was strictly applied in the first round. The parameters are estimated using data for the first round of offers by applying a regression with individual fixed effects. Production and price are significant explanatory variables in the third party treatment but not in the bargaining treatment, where the variables have very little explanatory power. The introduction of a third party induces the parties to make first offers that are strongly related to their relative production. Offers that are more strongly related to production will be closer to an equal split on average because the experiment is based on a random matching of pairs with a different production of words. This can explain the increased average share that person one offers in the first round in the third party treatment.

Table 2: Effect of relative production and price on first offers

	T I: Bargaining	T II: Third party
Production	-0.079	0.732***
	(0.363)	(0.083)
Price	0.047	0.063***
	(0.050)	(0.014)
Constant	0.538**	0.077*
	(0.215)	(0.046)
R^2	0.03	0.65
n	112	96/81

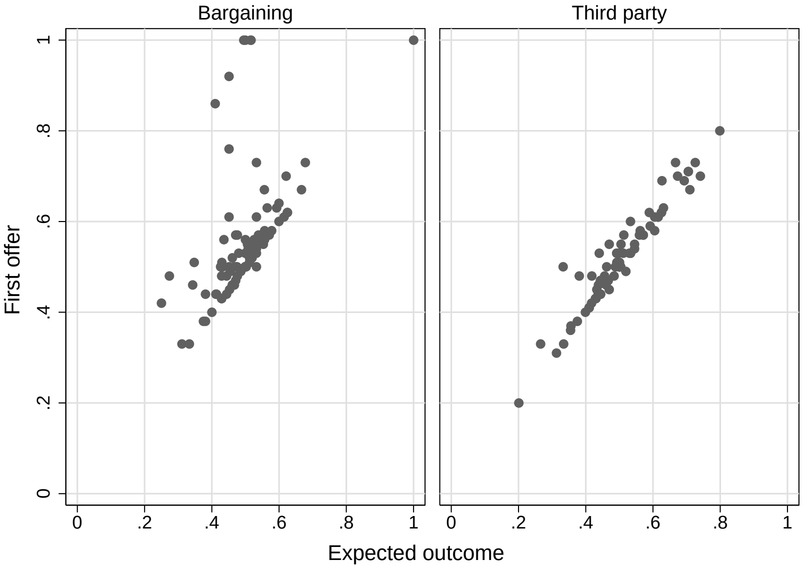
Notes. Standard errors in parentheses.***/**/*: 1/5/10% significance. Regression with individual fixed effects.

Figure 4 shows that there is a strong correlation between what a person expects will be the outcome from the negotiation and that person's first offer. However, in the bargaining treatment, there are a number of observations where negotiators start out by claiming for themselves a much higher share in the first offer than they expect will be the outcome, which is not the case for the third party treatment. The presence of a third party induces negotiators to make first offers that are more in line with expectations.

There are five first offers of zero to person B in the negotiation phase of the experiment. All these first offers were made by negotiators who also offered zero in the dictator game. This indicates that there is consistent behaviour over the different phases of the experiment. A dummy variable is constructed in order to check whether the dictator results have predictive power in explaining the negotiation efficiency. The dummy variable is zero if both of the dictator decisions are at the same level or higher (by a margin of at least 0.025) than the lowest offer that follows from the application of any of the principles described in Section 2; otherwise the dummy variable has a value of one. About one-third of the negotiators make offers in the dictator game that are lower than what the most self-serving principle tells them to do. This dummy variable is significantly correlated with both dispute costs and first offers ($p < 0.001$, Spearman rank-order correlation). It seems that inefficiency in negotiations is related to specific types of players.[10]

[10] Charness (2000) finds that different types of negotiators are important for explaining differ-

Figure 4: Correlation between expected outcome and first offer



4.2 Distributional properties of the outcome

The distributional properties of the settlements also differ between the two treatments, where there is more inequality in the third party treatment measured with the Gini index.[11] There is a significant increase in average inequality from 7% in the bargaining treatment to 12% in the third party treatment ($p = 0.008$, Wilcoxon rank-sum test). This excludes the 15 observations where a third party decided the settlement. If these observations are included, there is an even larger increase in the inequality.

The production phase of the experiment has explanatory power for the distributional properties of the negotiation outcome. In Table 3 the share that person A receives in the agreement is explained by his relative production in the pair, and his relative price in the pair. The effects are estimated by a regression with individual fixed effects. Both production and price are significant explanatory variables in the regressions. We see that the production coefficient is higher in the third party treatment, which indicates that there are more agreements closer to the proportionality principle in this treatment. However, a regression with an interaction term for production and treatment (the third column) shows

ences in efficiency. He investigates the influence of social preferences in a bargaining experiment with screening of participants into two groups based on their generosity in a dictator game (giving more or less than 30%). The sorting of participants into bargaining pairs reduces overall dispute costs, primarily because of a reduction in these costs when two generous types are paired together.

[11] The Gini index is zero for an equal split and one for a split where one player receives everything. The result is robust for other measures of inequality.

that the importance of relative production is not statistically significant between treatments.

Table 3: Effect of relative production and price on agreements

	TI Bargaining	TII Third party	TI & TII Negotiation
Production	0.571***	0.696***	0.542***
	(0.094)	(0.113)	(0.117)
Price	0.061***	0.048**	0.054***
	(0.013)	(0.019)	(0.011)
Prod. × treat.			0.156
			(0.147)
Constant	0.149***	0.102	0.136***
	(0.055)	(0.062)	(0.044)
R^2	0.418	0.45	0.439
n	112	96	208

Notes. Standard errors in parentheses.***/**/*: 1/5/10% significance
Regression with individual fixed effects.

Before person A chooses a first offer, he is asked about the amount of money he expects to end up with in the agreement. There is no significant difference between the treatments in the expected outcome. However, the inequality of the expected agreement in the bargaining treatment is 11%, significantly higher than the 7% for the actual agreements in that treatment ($p < 0.001$, Wilcoxon signed-rank test). It seems that the negotiators do not fully expect the more equal distribution in the bargaining treatment. For the third party treatment there is no statistically significant difference between the average expected outcome and the average actual outcome.

The dictator game results show that the participants to a large extent also chose to divide the production value according to each person's contribution when they had all the bargaining power. The coefficient estimate for production is 0.8 in Table 4, which indicates that offers are made close to the proportionality principle.[12]

[12] On average the participants offer 39% of the total production to the other player in the dictator game. The average offer to the other person in this dictator game is higher than typical dictator games where the average offer is about 20% of dictator endowment (Camerer, 2003). The higher average offer in a context with a real effort production phase is in line with the results of Cappelen, Hole, Sørensen, and Tungodden (2007).

Table 4: Effect of relative production and price on dictator decisions

	Dictator
Production	0.802***
	(0.135)
Price	0.052***
	(0.019)
Constant	0.150*
	(0.077)
R^2	0.141
n	208

Notes. Standard errors in parentheses.
***/**/*: 1/5/10% significance.

4.3 Third party behaviour

In a questionnaire, all of the six participants who acted as third parties preferred the proportionality principle. Four of them also answered that they would consider the strategy of the negotiators in their evaluation, saying they would allocate less to negotiators that were offering less than what was reasonable. Two arbitrators said that they would follow a rule that deducted all the dispute cost from the one who had offered less than what they perceived as fair, reasoning that he was responsible for the dispute costs. Seven of the 15 third party decisions were settled close to an application of a proportionality principle, and the other eight seemed not to follow a strict interpretation of a principle. All the third parties considered their role as an impartial third party to be to find a fair solution, and only one mentioned that a third party mechanism may foster faster decisions and improved efficiency.

The hypothesis that the third party mechanically compromises the final offers is not supported by the data. The final offers seem to have little direct influence; only in four cases are the settlements close to one of the parties' final offers. There are seven cases of third party settlements outside of the final offer from the negotiators.

Incorrect beliefs about the third party behaviour, for example excessive optimism about the outcome of a third party award, could cause the use of an expensive third party mechanism that would not have been used if the beliefs were correct (Babcock and Loewenstein, 1997). In the experiment, the participant who requests the third party is asked about what he believes will be the decision of the third party. As only 15 cases here were decided by a third party, we should be cautious interpreting this data, but the average belief about the third party decision is not significantly different from what the third party actually decides. More than half of the participants' subjective beliefs are within a 10% deviation from the third party decisions, and there is no evidence of systematic deviation of the beliefs about the third party decision.

4.4 Negotiators' answers to questionnaire

To understand what the participants think about fairness and bargaining, a short questionnaire was given to the negotiators after all the negotiations were completed. The first question given to the negotiators contains a brief description of a negotiation problem illustrated with three examples, similar to the actual cases that the participants experienced during the experiment. The answers show the largest support among the negotiators for the principle of *proportionality*, followed by the *laissez-faire* principle. Only 4% of the participants favour a strictly *egalitarian* division of the production value; see Table 5. These numbers seem to be biased by the actual experience during the experiment because the *proportionality* principle is favoured by 77% of those who participated in the third party treatment, and by only 57% of those who participated in the bargaining treatment, and there is a corresponding change in the support for the *laissez-faire* principle. The higher support for the *laissez-faire* principle in bargaining may reflect the self-serving use of such a principle during the bargaining treatment and a justification of this in the questionnaire.

Table 5: Preferred fairness principle

	I: Bargaining	II: Third party
Proportional to production	57%	77%
Laissez-faire principle	39%	19%
Equal division	4%	4%
n	56	48

The second question is related to whether they find it acceptable to use fairness or power in negotiations. The question is the same as that used by Binmore, Swierzbinski, Hsu, and Proulx (1993), who find that 35% say that one ought to play fair. Here, 57% of the negotiators in the bargaining treatment say that one ought to play fair compared with 69% of the negotiators in the third party treatment. The rest say that it is acceptable to use one's bargaining power. The experimental design differs from that of Binmore et al. (1993), and the larger support for fair play is probably because of the inclusion of a real effort production phase.

The third question relates to the use of bargaining strategies. Negotiators are asked to rank four alternative strategies according to what they think is the most important in negotiations in order to reach an agreement where they achieve their own goals. The results show that having a strong opening position is given the best overall rank; see Table 6. Although fairness is important in negotiations, players' views are more balanced when a fairness strategy is compared with other strategies.

Table 6: The importance of negotiation strategies

	Rank
A strong opening position	1
Seeking a fair outcome	2
More bargaining power	3
Willingness to make concessions	4
n	104

5 Related literature

This paper is related to several papers that study arbitration in an experimental setting (Ashenfelter et al., 1992; Bolton and Katok, 1998; Charness, 2000; Dickinson, 2004). They typically find that dispute rates more than double when conventional arbitration is introduced into negotiations. An innovative part of Ashenfelter et al. (1992) is the design where the arbitration decision is implemented as a computer random draw from a normal distribution with equal split of the outcome as the mean. A bargaining treatment is compared with an arbitration treatment. They measure the dispute rates as the number of negotiations where no agreement is reached after a fixed time period has elapsed. In the bargaining treatment, everything is lost after a certain time period, and this is compared to a forced arbitration settlement. The authors recognize that this experimental design implicitly raises the costs of no agreement compared with the treatment with forced arbitration because the likelihood of receiving zero from arbitration is very small.[13] This is different in the experimental design used here. The third party settlement is *not* enforced upon negotiators that do not close before a deadline, but it is a choice for negotiators to call upon a third party at any time during the negotiation. This implies that a negotiator can, if he believes that further negotiations would be costly, immediately submit the case to a third party and save negotiation costs.

Negotiation situations that facilitate the formation of different initial entitlements have been studied by Gächter and Riedl (2005, 2006). They find strong

[13] It is also the practice in these experiments to show the negotiators previous decisions by the arbitrator in the form of draws from a normal distribution. This information has potentially little value because there is nothing about the background history of offers or possible entitlement claims that the arbitrator would consider before a decision is reached. In the experiment reported here no information is provided about the arbitrators except that they are randomly drawn from among the participants in the room.

effects of entitlements on bargaining behaviour in an experiment where participants know whether they rank above or below the median answer to a general knowledge quiz. They find that most of the participants choose to split the endowment after a loss proportional to the entitlements that are suggested to them before the loss occurs. Gächter and Riedl (2006) find that proportionality is preferred in a questionnaire survey and that equality is more prevalent in actual negotiations.

The overall importance of entitlement in negotiations is confirmed in the experiment reported here. The literature has shown that proportionality is a strong principle held by many people in production contexts. The introduction of a third party settlement option induces the negotiators to change their strategy so that the offers are even more proportional to individual production. This paper adds to the literature by showing that a third party settlement option can increase the efficiency of bargaining in the sense of reducing the costs associated with rounds of alternating offers.

6 Concluding remarks

The costs associated with transactions in the broad sense constitute a large share of the economy, and the efficiency of institutions that facilitate transactions is of great importance for economic performance (North, 1990). Substantial resources are devoted to the formation and enforcement of contracts, and the resolution of disputes through arbitration, mediation and, of course, the legal system. It is important to build institutional arrangements that provide flexibility for people to negotiate their own solution, but at the same time provide efficient mechanisms to settle bargaining impasses.

The experimental results showed that both the negotiation efficiency and the distribution of payoffs are affected by the introduction of a third party. There is a significant reduction in dispute costs with the possibility of third party settlement. The introduction of a choice of a third party solution reduced dispute costs, primarily because it allows negotiators to cut short unfair treatment. The experiment provided an example of a third party mechanism that reduces the dispute costs, which is in contrast to the previous literature on arbitration. An implication for the efficiency of the design of a third party mechanism is that the option to submit the case to a third party should be available throughout the entire negotiation process, and not only after a period of time has elapsed. This ensures negotiators cut short unfair demands that would possibly lead to long and costly negotiations.

The efficiency gain from an option to use a third party is accompanied by an increase in the inequality of the distribution of gains, more in line with each person's contribution. The change in the distributional properties of the outcome that results from the introduction of a possible third party settlement raises a normative question of whether it is acceptable to influence the settlements such that other allocations are implemented than would result from negotiations between the parties without interference. In the experiment reported here, third parties favoured a fairness principle that was supported by a majority of negotiators. But the principle that a third party apply can run counter to the principle of fairness that has a broader legitimacy in society. Dworkin (2006) argues that legal theory at the adjudicative stage should require judges not only to uphold values of efficiency and coordination, but also to look to morality to decide the law. An important question then becomes the selection procedures of third parties.

Throughout the paper, we assumed that the parties agreed on the use of a specific third party mechanism. To agree on the use of such a mechanism during contracting is a negotiation in itself. It would therefore be interesting for further research to investigate how the commitment to the use of a third party in the contract phase influences the post-contractual negotiation behaviour under possible third party settlement.

References

Ashenfelter, Orley, Janet Currie, Henry S. Farber, and Matthew Spiegel (1992). "An experimental comparison of dispute rates in alternative arbitration systems", *Econometrica*, 60(6): 1407–1433.

Babcock, Linda and George Loewenstein (1997). "Explaining bargaining impasse: The role of self-serving biases", *Journal of Economic Perspectives*, 11(1): 109–126.

Bazerman, Max H. (1985). "Norms of distributive justice in interest arbitration", *Industrial and Labor Relations Review*, 38(4): 558–570.

Binmore, Ken, Joe Swierzbinski, Steven Hsu, and Chris Proulx (1993). "Focal points and bargaining", *International Journal of Game Theory*, 22: 381–409.

Bloom, David E. (1986). "Empirical models of arbitration behavior under conventional arbitration", *The Review of Economics and Statistics*, 68: 578–585.

Bolton, Gary E. and Elena Katok (1998). "Reinterpreting arbitration's narcotic effect: An experimental study of learning in repeated bargaining", *Games and Economic Behavior*, 25: 1–33.

Bruyn, Arnaud De and Gary E. Bolton (2008). "Estimating the influence of fairness on bargaining behavior", *Management Science*, 54(10): 1774–1791.

Camerer, Colin (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press.

Cappelen, Alexander W., Astrid D. Hole, Erik Ø. Sørensen, and Bertil Tungodden (2007). “The pluralism of fairness ideals: An experimental approach”, *American Economic Review*, 97(3): 818–827.

Charness, Gary (2000). “Bargaining efficiency and screening: an experimental investigation”, *Journal of Economic Behaviour and Organization*, 42: 285–304.

Dickinson, David L. (2004). “A comparison of conventional, final-offer, and ”combined” arbitration for dispute resolution”, *Industrial and Labor Relations Review*, 57(2): 288–301.

Dworkin, Ronald (2006). *Justice in Robes*, The Belknap Press of Harvard University Press.

Gächter, Simon and Arno Riedl (2005). “Moral property rights in bargaining with infeasible claims”, *Management Science*, 51(2): 249–263.

Gächter, Simon and Arno Riedl (2006). “Dividing justly in bargaining problems with claims”, *Social Choice and Welfare*, 27: 571–594.

Konow, James (1996). “A positive theory of economic fairness”, *Journal of Economic Behaviour and Organisation*, 31: 13–35.

Konow, James (2000). “Fair shares: Accountability and cognitive dissonance in allocation decisions”, *The American Economic Review*, 90(4): 1072–1091.

North, Douglass C. (1990). *Institutions, Institutional Change and Economic Performance*, Cambridge University Press.

Ochs, Jack and Alvin E. Roth (1989). “An experimental study of sequential bargaining”, *American Economic Review*, 79(3): 355–384.

Rubinstein, Ariel (1982). “Perfect equilibrium in a bargaining model”, *Econometrica*, 50: 97–109.

Weg, Eythan, Amnon Rapoport, and Dan S. Felsenthal (1990). “Two-person bargaining behavior in fixed discounting factors game with infinite horizon”, *Games and Economic Behavior*, 2: 76–95.

Appendix A Instructions

This supplement contains all the instructions read to the participants during the experiment. The four subsections follow the structure of the experiment explained in Section 2 of the paper. This supplement contains instructions for both treatments. For the bargaining treatment, the instructions can be read by using all the text in square brackets and deleting all the text in curly brackets, and vice versa for the third party treatment. The instructions were given in Norwegian. The translation is by the author.

General introduction

Welcome to this experiment. My name is (...) and I will guide you through the experiment. The results from the experiment will be used in a research project, and it is therefore important that you all stick to the rules that have been distributed:

- You should not talk to other participants.
- If you have questions or problems during the experiment, raise your hand and we will come to you.
- You should not open other web pages.

If you breach these rules, you will have to leave the room. There will be pauses during the experiment and it is important that you sit still and keep quiet during these.

You will be completely anonymous in the experiment. You will not at any time be asked about who you are. It will not be possible for us or the other participants to find out which choices you have made. You will be asked to make choices in several different situations in this experiment. For every situation, you will be randomly connected to another person in this room. Your actual payment will be determined as follows: we randomly draw one of the situations you were involved in and pay the amount of money you received in that situation. The choices that you make will not influence which situation is drawn; it will be an entirely random draw and there is an equal chance for all situations to be drawn. You should therefore think about each situation as if it is the one that determines how much you earn.

When the experiment is finished, you will see a payment code on the screen. You are asked to write down this code on a form that will be sent to the accounting department at (...). Employees at the accounting department will receive a list of codes and amounts from us and match these with the payment instructions from the forms. This is done so that nobody will know how much you earned.

{There are two different roles in the experiment. Most of you are participants in negotiations, while some of you are randomly drawn to act as a third party. The content of the folder shows which role you have been assigned. Negotiators have only received text marked A or text marked B. Those of you that are third parties have, in addition to the two pieces of text, received in the folder at your desk a booklet labelled 'Examples and questions'.}

The experiment consists of four phases. I will now explain the main features of the experiment. I will stop before we start a new phase and explain in more detail what you should do in each phase. In the first phase of the experiment, you should copy text for 10 minutes in Word. You will be paid a price for each correct word you have typed. In phase two of the experiment you will be randomly matched with other persons in this room, and each of you in a pair will choose how much of the combined production value you will distribute to yourself and to the other person. You will be involved in four such *situations of distribution.*

In the third phase of the experiment, you will also be randomly matched with people in this room. You will then *negotiate* about the division of the combined production value by sending proposals to each other until one of you accepts the other's proposal. The production value shrinks by 4% every time one of you does not accept the other's proposal. {You will also have the opportunity to let a third person in this room decide the distribution. The third party does not have any other tasks than to decide on distributions that are sent to him or her.} You will be involved in four such *situations of negotiation.* In the last phase of the experiment you are asked to answer a few questions about the type of situations that you have experienced.

Introduction phase 1

The first thing [you] {negotiators} will do is to copy text from an official report that is marked with either an A or a B, and which you will find in the folder on your desk. You will copy the text into Word when I tell you to start. I will tell you when 10 minutes have passed and everybody must then stop. You will be paid for each correct word you type. You may use the spellchecker in Word.

{Some of you will be drawn to act as a third party and you will make decisions in particular situations if other participants request this later in the experiment. Those of you who are third parties should first read through the four examples of negotiation that are in the folder, and then answer the five questions in the folder. The answers should be written in Word at the same time the others copy text. The third parties will later be asked to provide answers to the questions on their screen. Those who are third parties will receive 300 kroner for the job. This amount is fixed and is not influenced by what you do in the experiment.}

{To everybody,} I remind you that you should raise your hand if you have any problems or questions, and then one of us in the research group will come and help you. You can now open a new document in Word and we will soon start to type. *You can start typing now .*

Everybody must now stop typing. [You] {Those of you that are negotiators} should now highlight all the text typed and copy it to the window in the Mozilla browser, then click on the button marked 'submit text'. {Those of you who are third parties should not do anything now; however, you will later be asked to submit your answers to the questions.}

After having submited the text you will see a screen that shows how much you have produced and the value of your production. The production is rounded off to the nearest 50 words. Half of [you] {negotiators} have copied text marked A, which is an excerpt from an official report on the merger of the telecoms, IT, and media sectors. You will receive one krone and 50 oere for each correct word you have typed. The other half of [you] {negotiators} have copied text marked B, which is an excerpt from an official report about Norwegian performing art. You will receive 75 oere for each correct word you have typed. These prices are randomly determined by us. Finally, click on the button marked ‘continue’.

Introduction phase 2

[You] {Negotiators} will now be randomly matched with other people in this room. In each *situation of distribution* you will not know who the other person is, and the other person will not know who you are. You will be informed about how many words he or she has produced and what price each of you has randomly been allocated. You will then choose a *distribution* of the combined production value between you and the other person. Remember that this is real money and the way that you divide the money determines how much you earn and how much the other person earns. You will be asked to make decisions in two such *situations of distribution.* In two other situations of distribution, another person will decide how much he or she will distribute to you.

After you have registered the distribution, you will see a new screen where you are asked either to confirm the distribution or to go back and change the distribution. When you have confirmed your choices, you will receive a message that you have finished the second phase of the experiment. You should then quietly wait for all the other people in the room to finish making choices in their situations. On the computer you will soon see a screen with the first situation and you can then start making choices. {Third parties can continue and will later be asked to deliver the answers.}

Introduction phase 3

Everybody has finished the second phase and I shall now explain what you will do in the third phase of the experiment. [You] {Negotiators} will this time also be randomly matched with other people in this room. In each *situation of negotiation* you will not know who the other person is and the other person will not know who you are. You will be informed about how many words the other person has produced and what price he or she has randomly been allocated. One of you is randomly drawn to make the first proposal for division of the combined production value. The proposal will be sent to the other person and he or she has [two] {three} choices: to accept your proposal or to make a new proposal for division {, or to decide that the distribution will be determined by a third person in the room}. New proposals are sent back and forth until one of you chooses to accept the proposal {or to give a third party the task of deciding the distribution}. Every time one of you does not accept the proposal for division, but comes up with a new proposal, the remaining production value will be reduced by 4%. {If you choose to let a third party decide the distribution, the remaining production value is reduced by another 6%, in total 10%. Third parties will have information about both participants’ production and negotiation history.} Everybody will be involved in four such situations of negotiation.

In some situations you will be asked what you think will be the final outcome of the negotiation {and how a possible third party will divide the amount}. If your answer is within a deviation of plus or minus 20 kroner of the actual result, you will receive 20 kroner in extra payment, with one exception: if you guessed the negotiation result in a situation, and this particular result was randomly drawn, you will *not* receive the extra payment for a correct guess but only the payout in this situation. You will also be asked to state how *certain* you are about your guess. Your answer should be given in terms of a certainty percentage, that is, a number between 0 and 100. It is important that you write a high percentage if you are certain that this will be the result, and a low percentage if you are uncertain if this will be the final result.

{Those of you that have been drawn to be a third party will within the next 10 minutes copy the answers to the questions from Word into the window on the browser. You will also paste your proposal for division of the four examples from the folder. When you have done that, click on 'submit answer' and you will be asked to wait until possible situations arise in which you are asked to decide. After a short period of time you will see these situations on the screen and you can then start to decide on the distributions.}

On the computer [you] {negotiators} will soon see a new screen with the first situation and you can then start to negotiate. When the situation is accepted {or one of you has chosen to have a third party decide the situation}, you will automatically get a new situation to negotiate. When you have finished all the negotiation situations, you will be asked to wait until everybody has finished their choices.

Introduction phase 4

Everybody has finished and we will soon draw the situation that will decide your payment from this experiment. First, we ask you to answer a few questions. Soon you will see a new screen with information about the first question. You should click on the button marked 'go forward' when you have read the information and thereafter you should answer all the questions.

Closing and payment

Everybody has now answered the questions. You will soon see a screen that informs you about which situation has randomly been drawn, and how much you earned in this situation. This screen will be open for 45 seconds. Thereafter you will automatically be forwarded to a new screen, which only contains a payment code.

Everybody now has a screen with the payment code. Write down this payment code on the form that you find in the folder next to you. On the form also write down your name and bank account details. Put the form in the envelope and place it in the box by the door when you leave the room.

The experiment is now finished and, on behalf of the research team, I thank you again for your participation in this experiment.

Appendix B Questionnaire

This supplement contains questions given to negotiators after the experiment had ended. The questionnaire was given in Norwegian. The translation is by the author.

Information

Person A and person B have produced a good or a service with a total value of 1000 kroner. The value of A's and B's production is determined by how much effort they have exerted and the price that they receive for what they have produced. The individual effort is determined both by how hard each has worked and by the individuals' skill in this type of work. The price is randomly determined and cannot be influenced by the individual.

Below you can see an example of such a situation. Click on the button below to see more examples. You will thereafter be asked to state what *you* think is the fairest way of dividing the combined production value. Note that the combined production value in all of the examples is 1000 kroner.

Production	**You**	**The other person**
Effort	200 units	200 units
Price of effort	4 kr	1 kr
Production value	800 kr	200 kr

Table 7: Example 1

Production	**You**	**The other person**
Effort	800 units	200 units
Price of effort	1 kr	1 kr
Production value	800 kr	200 kr

Table 8: Example 2

Production	**You**	**The other person**
Effort	150 units	400 units
Price of effort	4 kr	1 kr
Production value	600 kr	400 kr

Table 9: Example 3

Question 1

Now mark the principle that *you* think gives the fairest division in these types of situations. How would you divide the total production value between you and the other person?

○ Divide equally

○ Divide proportional to individual production value

○ Divide proportional to individual effort

Question 2

Is this the sort of situation in which people ought to play fair or is it socially acceptable to use whatever bargaining power one has?

○ Use bargaining power

○ Play fair

Question 3

Rank from 1 (very important) to 4 (not so important) what *you* believe is important in order to reach an agreement where you achieve your own goals in negotiations. Write a number from 1 to 4 in all the boxes below. You cannot write the same number in more than one box.

□ You are willing to make concessions

□ You have more bargaining power

□ You seek a fair outcome

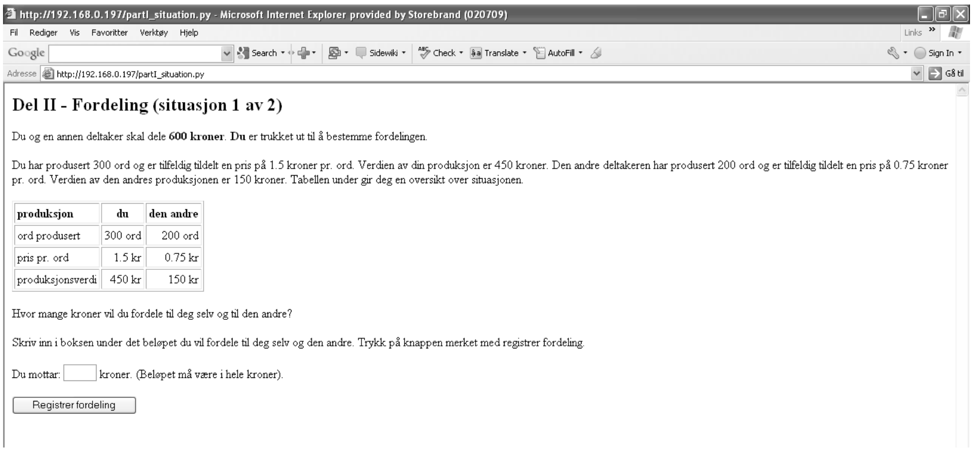
□ You have a strong opening position

Questions for third parties only

1. What do you think explains why a person would let a third party decide the outcome of a negotiation?
2. Which aspects of these negotiations would you emphasize as a third party? (*refers to four detailed examples on negotiations that are presented in writing*)
3. Why would you emphasize these aspects?
4. How do you think the possibility of letting a third party decide the outcome of a negotiation influences the proposals made by the parties during a negotiation?
5. Will the proposals that the parties have made during the negotiation influence your judgement as a third party? If so, how?

Appendix C Selected screenshots

This supplement contains a selection of the screenshots from the experiment. In the heading of the screenshot there is a reference to the phase of the experiment explained in Section 2 of the paper. This supplement contains screenshots from both treatments. The screenshots are in Norwegian. The translation of the main text on the screen follows below the screenshots. The translation is by the author.


http://192.168.0.197/partI_situation.py - Microsoft Internet Explorer provided by Storebrand (020709)
Fil Rediger Vis Favoritter Verktøy Hjelp
Links
Google
Search
Sidewiki
Check
Translate
AutoFill
Sign In
Adresse http://192.168.0.197/partI_situation.py
Gå til

Del II - Fordeling (situasjon 1 av 2)

Du og en annen deltaker skal dele **600 kroner**. **Du** er trukket ut til å bestemme fordelingen.

Du har produsert 300 ord og er tilfeldig tildelt en pris på 1.5 kroner pr. ord. Verdien av din produksjon er 450 kroner. Den andre deltakeren har produsert 200 ord og er tilfeldig tildelt en pris på 0.75 kroner pr. ord. Verdien av den andres produksjonen er 150 kroner. Tabellen under gir deg en oversikt over situasjonen.

produksjon	**du**	**den andre**
ord produsert	300 ord	200 ord
pris pr. ord	1.5 kr	0.75 kr
produksjonsverdi	450 kr	150 kr

Hvor mange kroner vil du fordele til deg selv og til den andre?

Skriv inn i boksen under det beløpet du vil fordele til deg selv og den andre. Trykk på knappen merket med registrer fordeling.

Du mottar: □ kroner. (Beløpet må være i hele kroner).

Registrer fordeling

Figure 5: Dictator game

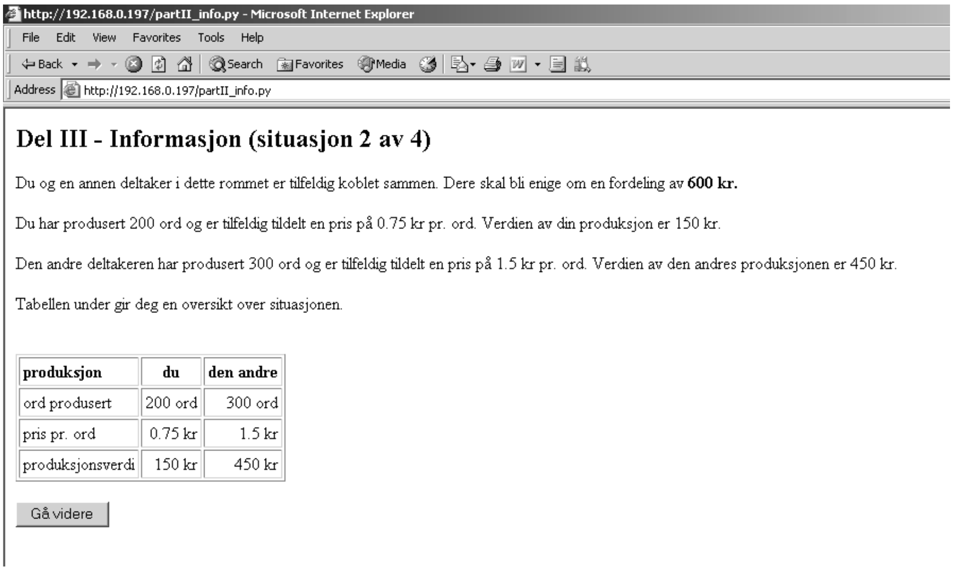
Phase II - Distribution (situation 1 of 2)

You and another participant shall divide **600 kroner**. **You** have been drawn to decide the distribution.
You have produced 300 words and are randomly assigned a price of 1.5 kroner per word. The value of your production is 450 kroner. The other participant has produced 200 words and is randomly assigned a price of 0.75 kroner per word. The value of the other's production is 150 kroner. The table below gives you an overview of the situation.

production	**you**	**the other**
words produced	300 words	200 words
price per word	1.5 kr	0.75 kr
production value	450 kr	150 kr

How many kroner will you distribute to yourself and to the other?
Write into the box below the amount you will distribute to yourself and the other. Click on the button marked with submit distribution.
You receive: □ kroner. (The amount must be in whole kroner).

Figure 6: Information about a bargaining situation


http://192.168.0.197/partII_info.py - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://192.168.0.197/partII_info.py
Gå videre

Del III - Informasjon (situasjon 2 av 4)

Du og en annen deltaker i dette rommet er tilfeldig koblet sammen. Dere skal bli enige om en fordeling av **600 kr.**

Du har produsert 200 ord og er tilfeldig tildelt en pris på 0.75 kr pr. ord. Verdien av din produksjon er 150 kr.

Den andre deltakeren har produsert 300 ord og er tilfeldig tildelt en pris på 1.5 kr pr. ord. Verdien av den andres produksjonen er 450 kr.

Tabellen under gir deg en oversikt over situasjonen.

produksjon	**du**	**den andre**
ord produsert	200 ord	300 ord
pris pr. ord	0.75 kr	1.5 kr
produksjonsverdi	150 kr	450 kr

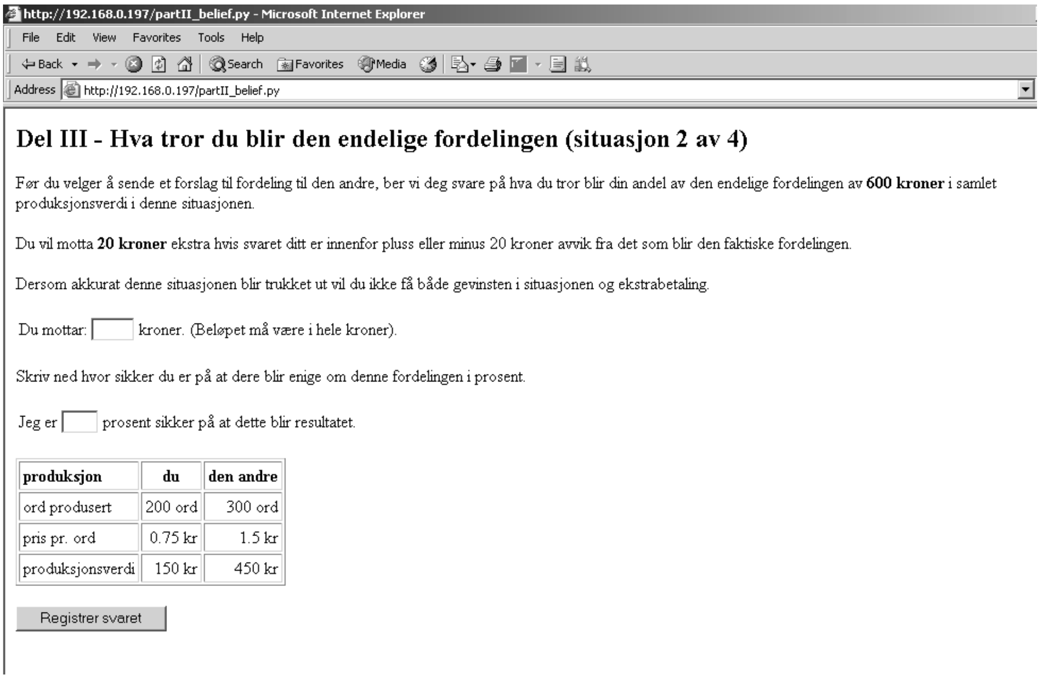
Phase III - Information (situation 2 of 4)

You and another participant in this room have been randomly matched. You shall agree about the division of **600 kr**.

You have produced 200 words and are randomly assigned a price of 0.75 kr per word. The value of your production is 150 kr. The other participant has produced 300 words and is randomly assigned a price of 1.5 kr per word. The value of the other's production is 450 kr. The table below gives you an overview of the situation.

production	**you**	**the other**
words produced	200 words	300 words
price per word	0.75 kr	1.5 kr
production value	150 kr	450 kr

Figure 7: Beliefs about the outcome of the bargaining


http://192.168.0.197/partII_belief.py - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://192.168.0.197/partII_belief.py
Registrer svaret

Del III - Hva tror du blir den endelige fordelingen (situasjon 2 av 4)

Før du velger å sende et forslag til fordeling til den andre, ber vi deg svare på hva du tror blir din andel av den endelige fordelingen av **600 kroner** i samlet produksjonsverdi i denne situasjonen.

Du vil motta **20 kroner** ekstra hvis svaret ditt er innenfor pluss eller minus 20 kroner avvik fra det som blir den faktiske fordelingen.

Dersom akkurat denne situasjonen blir trukket ut vil du ikke få både gevinsten i situasjonen og ekstrabetaling.

Du mottar: □ kroner. (Beløpet må være i hele kroner).

Skriv ned hvor sikker du er på at dere blir enige om denne fordelingen i prosent.

Jeg er □ prosent sikker på at dette blir resultatet.

produksjon	**du**	**den andre**
ord produsert	200 ord	300 ord
pris pr. ord	0.75 kr	1.5 kr
produksjonsverdi	150 kr	450 kr

Phase III - What do you think will be the finale division (situation 2 of 4)

Before you choose to send a proposal for division to the other, we ask you to answer what you think will be your share the final division of **600 kroner** in total production value in this situation.

You will receive **20 kroner** extra if your answer is within plus or minus 20 kroner deviation from what the actual division turns out to be. If this particular situation is drawn you will not receive both the earnings in the situation and the extra earnings.

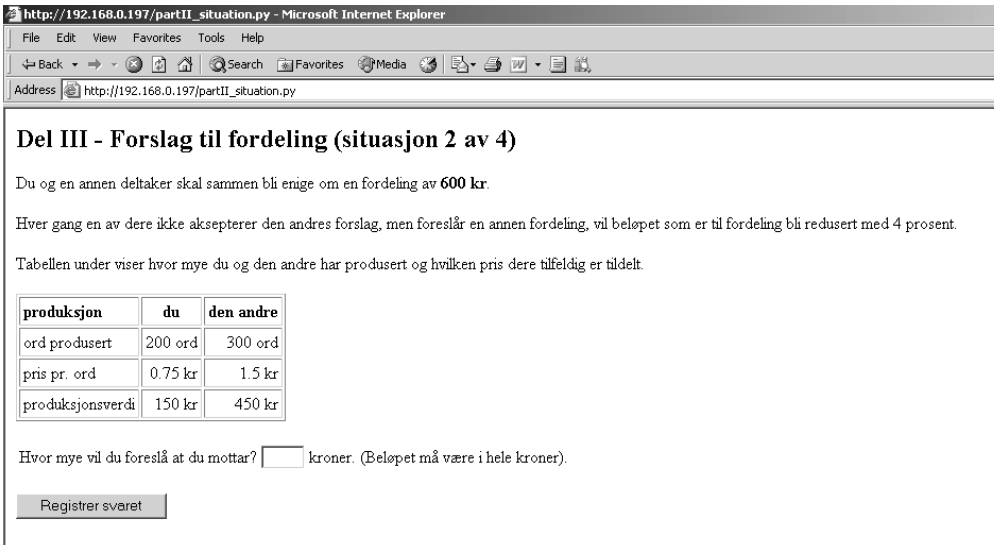
You receive: □ kroner. (The amount must be in whole kroner).

Write down in per cent how sure you are that you will agree about this division.

I am □ per cent sure that this will be the result.

production	**you**	**the other**
words produced	200 words	300 words
price per word	0.75 kr	1.5 kr
production value	150 kr	450 kr

Figure 8: The first offer decision



http://192.168.0.197/partII_situation.py - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://192.168.0.197/partII_situation.py

Del III - Forslag til fordeling (situasjon 2 av 4)

Du og en annen deltaker skal sammen bli enige om en fordeling av **600 kr**.

Hver gang en av dere ikke aksepterer den andres forslag, men foreslår en annen fordeling, vil beløpet som er til fordeling bli redusert med 4 prosent.

Tabellen under viser hvor mye du og den andre har produsert og hvilken pris dere tilfeldig er tildelt.

produksjon	du	den andre
ord produsert	200 ord	300 ord
pris pr. ord	0.75 kr	1.5 kr
produksjonsverdi	150 kr	450 kr

Hvor mye vil du foreslå at du mottar? □ kroner. (Beløpet må være i hele kroner).

Registrer svaret

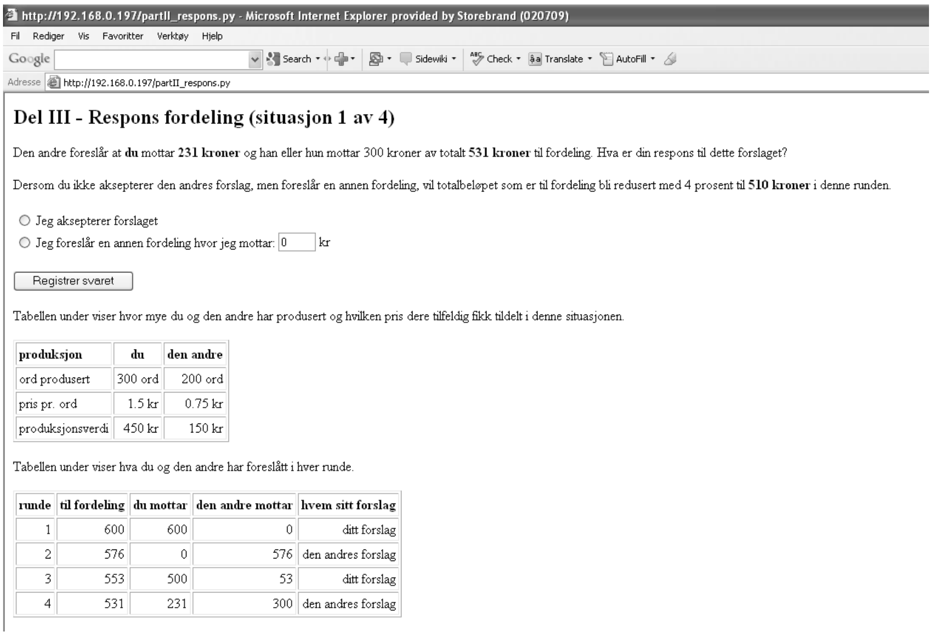
Phase III - Proposal to division (situation 2 of 4)

You and another participant shall together agree on a division of **600 kr**.
Every time one of you does not accept the others proposal, but suggests a different division, the total amount to divide will be reduced by 4 per cent.
The table below shows how much you and the other have produced and which prices that you have been randomly assigned.

production	you	the other
words produced	200 words	300 words
price per word	0.75 kr	1.5 kr
production value	150 kr	450 kr

How much do you propose that you receive? □ kroner. (The amount must be in whole kroner).

Figure 9: The reoffer decision in the bargaining treatment



http://192.168.0.197/partII_respons.py - Microsoft Internet Explorer provided by Storebrand (020709)
Fil Rediger Vis Favoritter Verktøy Hjelp
Google Search Sidewiki Check Translate AutoFill
Adresse http://192.168.0.197/partII_respons.py

Del III - Respons fordeling (situasjon 1 av 4)

Den andre foreslår at **du** mottar **231 kroner** og han eller hun mottar 300 kroner av totalt **531 kroner** til fordeling. Hva er din respons til dette forslaget?

Dersom du ikke aksepterer den andres forslag, men foreslår en annen fordeling, vil totalbeløpet som er til fordeling bli redusert med 4 prosent til **510 kroner** i denne runden.

○ Jeg aksepterer forslaget
○ Jeg foreslår en annen fordeling hvor jeg mottar: 0 kr

Registrer svaret

Tabellen under viser hvor mye du og den andre har produsert og hvilken pris dere tilfeldig fikk tildelt i denne situasjonen.

produksjon	du	den andre
ord produsert	300 ord	200 ord
pris pr. ord	1.5 kr	0.75 kr
produksjonsverdi	450 kr	150 kr

Tabellen under viser hva du og den andre har foreslått i hver runde.

runde	til fordeling	du mottar	den andre mottar	hvem sitt forslag
1	600	600	0	ditt forslag
2	576	0	576	den andres forslag
3	553	500	53	ditt forslag
4	531	231	300	den andres forslag

Phase III - Response division (situation 1 of 4)

The other proposes that **you** receive **231 kroner** and that he or she receives 300 kroner of the total of **531 kroner** to divide. What is your response to this proposal?
If you do not accept the other's proposal, but make a different proposal, the total amount to divide will be reduced by 4 per cent to **510 kroner** in this round.

○ I accept the proposal
○ I propose a different division where I receive □ kr.

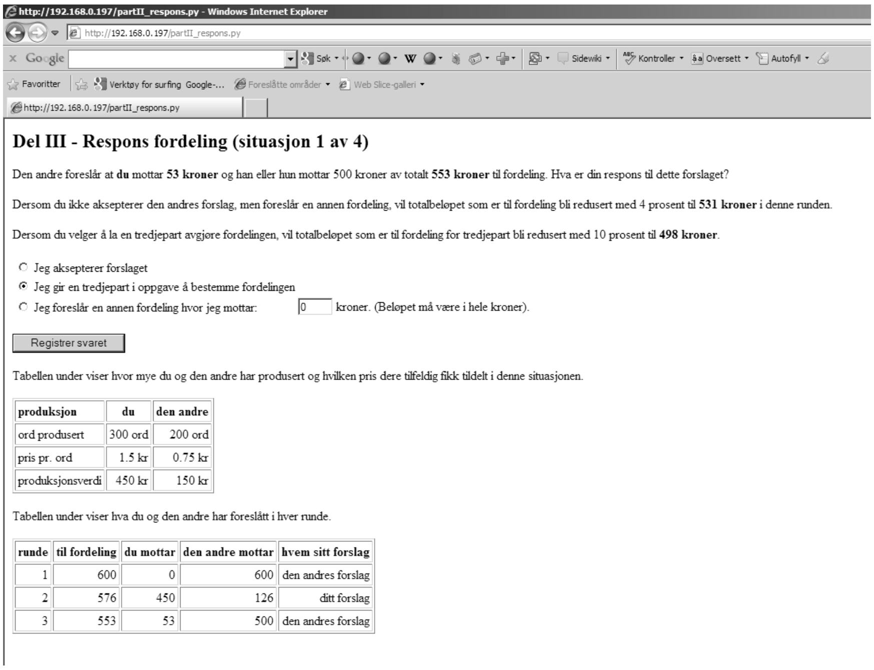
The table below shows how much you and the other have produced and which prices that you have been randomly assigned in this situation.

production	you	the other
words produced	300 words	200 words
price per word	1.5 kr	0.75 kr
production value	450 kr	150 kr

The table below shows what you and the other have proposed in each round.

round	to divide	you receive	the other receives	who made the proposal
1	600	600	0	your proposal
2	576	0	576	the other's proposal
3	553	500	53	your proposal
4	531	231	300	the other's proposal

Figure 10: The reoffer decision in the third party treatment


http://192.168.0.197/partII_respons.py - Windows Internet Explorer
http://192.168.0.197/partII_respons.py
Google
Søk
Sidewiki
Kontroller
Oversett
Autofyll
Favoritter
Verktøy for surfing Google-...
Foreslåtte områder
Web Slice-galleri
http://192.168.0.197/partII_respons.py

Del III - Respons fordeling (situasjon 1 av 4)

Den andre foreslår at **du** mottar **53 kroner** og han eller hun mottar 500 kroner av totalt **553 kroner** til fordeling. Hva er din respons til dette forslaget?

Dersom du ikke aksepterer den andres forslag, men foreslår en annen fordeling, vil totalbeløpet som er til fordeling bli redusert med 4 prosent til **531 kroner** i denne runden.

Dersom du velger å la en tredjepart avgjøre fordelingen, vil totalbeløpet som er til fordeling for tredjepart bli redusert med 10 prosent til **498 kroner**.

○ Jeg aksepterer forslaget
● Jeg gir en tredjepart i oppgave å bestemme fordelingen
○ Jeg foreslår en annen fordeling hvor jeg mottar: 0 kroner. (Beløpet må være i hele kroner).

Registrer svaret

Tabellen under viser hvor mye du og den andre har produsert og hvilken pris dere tilfeldig fikk tildelt i denne situasjonen.

produksjon	du	den andre
ord produsert	300 ord	200 ord
pris pr. ord	1.5 kr	0.75 kr
produksjonsverdi	450 kr	150 kr

Tabellen under viser hva du og den andre har foreslått i hver runde.

runde	til fordeling	du mottar	den andre mottar	hvem sitt forslag
1	600	0	600	den andres forslag
2	576	450	126	ditt forslag
3	553	53	500	den andres forslag

Phase III - Response division (situation 1 of 4)

The other proposes that **you** receive **53 kroner** and that he or she receives 500 kroner of the total of **553 kroner** to divide. What is your response to this proposal?
If you do not accept the other's proposal, but make a different proposal, the total amount to divide will be reduced by 4 per cent to **531 kroner** in this round.
If you choose to let a third party decide the division, the total amount that the third party shall divide will be reduced by 10 per cent to **498 kroner**.

○ I accept the proposal
○ I give a third party the task of deciding the division
○ I propose a different division where I receive □ kroner.

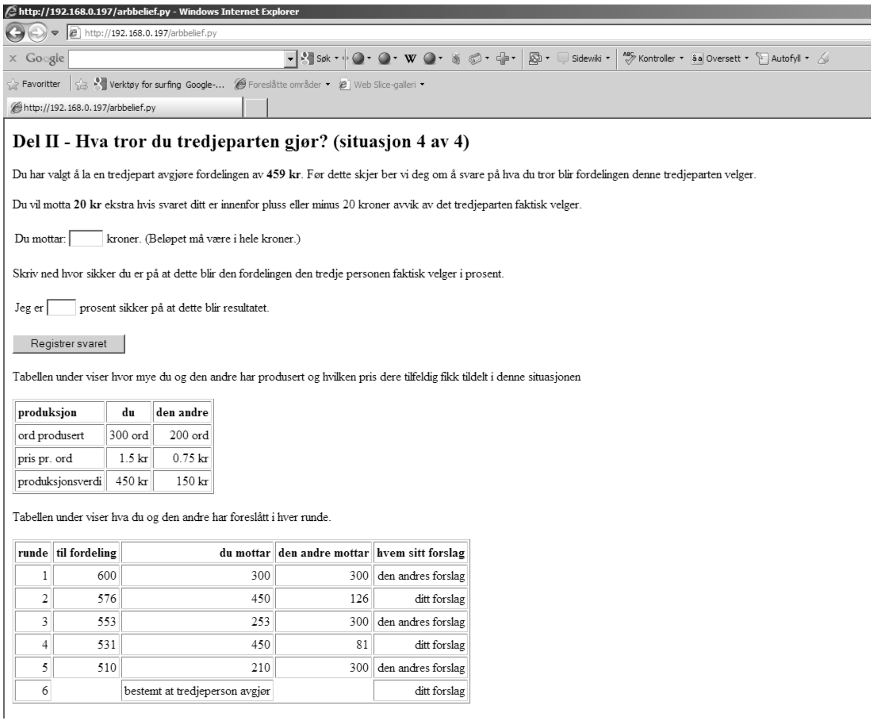
The table below shows how much you and the other have produced and which prices that you have been randomly assigned in this situation.

production	you	the other
words produced	300 words	200 words
price per word	1.5 kr	0.75 kr
production value	450 kr	150 kr

The table below shows what you and the other have proposed in each round.

round	to divide	you receive	the other receives	who made the proposal
1	600	0	600	the other's proposal
2	576	450	126	your proposal
3	553	53	500	the other's proposal

Figure 11: Beliefs about the third party decision


http://192.168.0.197/arbbelief.py - Windows Internet Explorer
http://192.168.0.197/arbbelief.py
Google
Søk
Sidewiki
Kontroller
Oversett
Autofyll
Favoritter
Verktøy for surfing Google-...
Foreslåtte områder
Web Slice-galleri
http://192.168.0.197/arbbelief.py

Del II - Hva tror du tredjeparten gjør? (situasjon 4 av 4)

Du har valgt å la en tredjepart avgjøre fordelingen av **459 kr**. Før dette skjer ber vi deg om å svare på hva du tror blir fordelingen denne tredjeparten velger.

Du vil motta **20 kr** ekstra hvis svaret ditt er innenfor pluss eller minus 20 kroner avvik av det tredjeparten faktisk velger.

Du mottar: □ kroner. (Beløpet må være i hele kroner.)

Skriv ned hvor sikker du er på at dette blir den fordelingen den tredje personen faktisk velger i prosent.

Jeg er □ prosent sikker på at dette blir resultatet.

Registrer svaret

Tabellen under viser hvor mye du og den andre har produsert og hvilken pris dere tilfeldig fikk tildelt i denne situasjonen

produksjon	du	den andre
ord produsert	300 ord	200 ord
pris pr. ord	1.5 kr	0.75 kr
produksjonsverdi	450 kr	150 kr

Tabellen under viser hva du og den andre har foreslått i hver runde.

runde	til fordeling	du mottar	den andre mottar	hvem sitt forslag
1	600	300	300	den andres forslag
2	576	450	126	ditt forslag
3	553	253	300	den andres forslag
4	531	450	81	ditt forslag
5	510	210	300	den andres forslag
6		bestemt at tredjeperson avgjør		ditt forslag

Phase III - What do you think the third party will do? (situation 4 of 4)

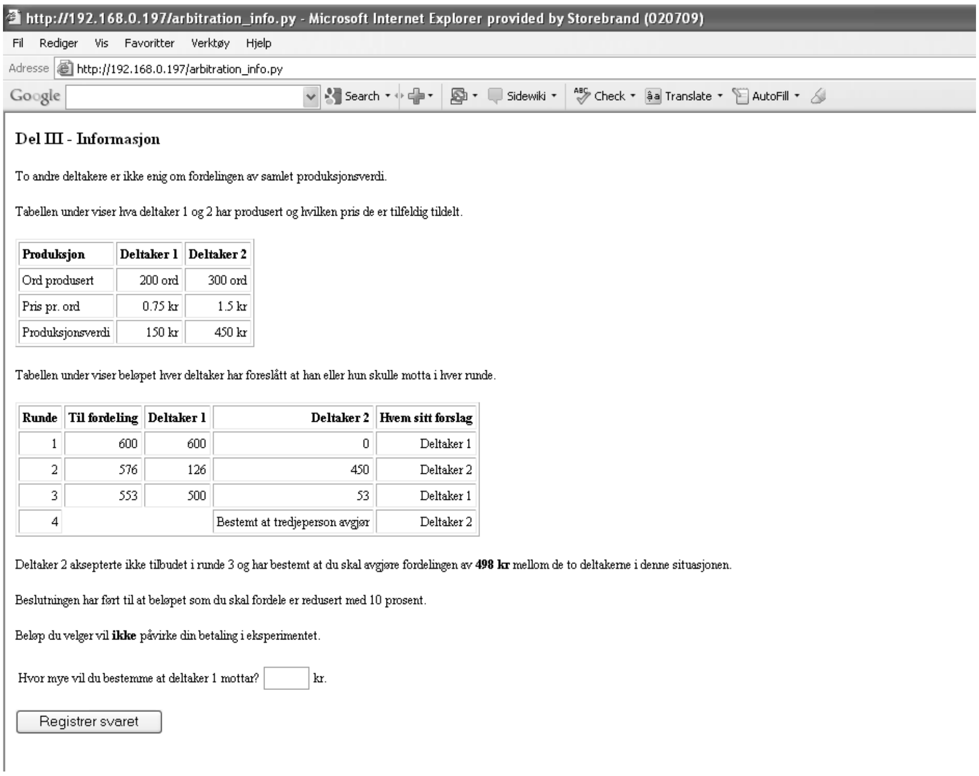
You have chosen to let a third party decide the division of **459 kr**. Before that happens we ask you to answer what you think will be the division that this third party chooses.
You will receive **20 kroner** extra if your answer is within plus or minus 20 kroner deviation from what the third party actually decide.
You receive: □ kroner. (The amount must be in whole kroner).
Write down in per cent how sure you are that this is the division that the third party actually chooses.
I am □ per cent sure that this will be the result.
The table below shows how much you and the other have produced and which prices that you have been randomly assigned in this situation.

production	you	the other
words produced	300 words	200 words
price per word	1.5 kr	0.75 kr
production value	450 kr	150 kr

The table below shows what you and the other have proposed in each round.

round	to divide	you receive	the other receives	who made the proposal
1	600	300	300	the other's proposal
2	576	450	126	your proposal
3	553	253	300	the other's proposal
4	531	450	81	your proposal
5	510	210	300	the other's proposal
6		Let third party decide		your proposal

Figure 12: The third party decision


http://192.168.0.197/arbitration_info.py - Microsoft Internet Explorer provided by Storebrand (020709)
Fil Rediger Vis Favoritter Verktøy Hjelp
Adresse http://192.168.0.197/arbitration_info.py
Google Search Sidewiki Check Translate AutoFill

Del III - Informasjon

To andre deltakere er ikke enig om fordelingen av samlet produksjonsverdi.

Tabellen under viser hva deltaker 1 og 2 har produsert og hvilken pris de er tilfeldig tildelt.

Produksjon	Deltaker 1	Deltaker 2
Ord produsert	200 ord	300 ord
Pris pr. ord	0.75 kr	1.5 kr
Produksjonsverdi	150 kr	450 kr

Tabellen under viser beløpet hver deltaker har foreslått at han eller hun skulle motta i hver runde.

Runde	Til fordeling	Deltaker 1	Deltaker 2	Hvem sitt forslag
1	600	600	0	Deltaker 1
2	576	126	450	Deltaker 2
3	553	500	53	Deltaker 1
4			Bestemt at tredjeperson avgjør	Deltaker 2

Deltaker 2 aksepterte ikke tilbudet i runde 3 og har bestemt at du skal avgjøre fordelingen av **498 kr** mellom de to deltakerne i denne situasjonen.

Beslutningen har ført til at beløpet som du skal fordele er redusert med 10 prosent.

Beløp du velger vil **ikke** påvirke din betaling i eksperimentet.

Hvor mye vil du bestemme at deltaker 1 mottar? □ kr.

Registrer svaret

Phase III - Information

Two other participants have not agreed on the division of a total production value.
The table below shows what participants 1 and 2 have produced and which prices they have randomly been assigned.

Production	Participant 1	Participant 2
Words produced	200 words	300 words
Price per word	0.75 kr	1.5 kr
Production value	150 kr	450 kr

The table below shows the amounts that each participant has proposed that he or she should receive in each round.

Round	To divide	Participant 1	Participant 2	who made the proposal
1	600	600	0	Participant 1
2	576	126	450	Participant 2
3	553	500	53	Participant 1
4			Let third party decide	Participant 2

Participant 2 did not accept the offer in round 3 and has decided that you shall decide the division of **498 kr** between the two participants in this situation.
The decision has reduced the amount that you shall divide with 10 per cent.
The amount you choose will **not** affect your own payment in the experiment.
How much do you decide that participant 1 receives? □ kr.

Fairness motivation in bargaining[*]

Sigbjørn Birkeland d.y.[†]
Department of Economics
Norwegian School of Economics and Business Administration

10th of April

Abstract

In this paper, we develop a model that captures the potential conflict between two individuals who follow different fairness principles in bargaining. This model is used to analyse the influence of fairness motivation on the possibility of reaching an agreement in bargaining, and to examine the properties of the agreement. We show that bargaining between two individuals who are strongly fairness motivated, but who disagree about what represents a fair division, ends in disagreement. This result contrasts the standard bargaining model with individuals who are only motivated by material self-interest, which always leads to agreement. Furthermore, by applying the Nash bargaining solution, we study the influence of fairness motivation on the bargaining outcome. A fairness motivated individual reaches an outcome that is closer to his fairness principle in bargaining against an individual who is only motivated by material self-interest.

[*] I would very much like to thank Bertil Tungodden, Alexander Cappelen, James Konow, and Shachar Kariv for valuable comments and helpful discussions. Thanks also to participants at the 5th Nordic Conference on Behavioural and Experimental Economics at Aalto University.

[†] Address: Helleveien 30, 5045 Bergen, Norway. E-mail sigbjorn.birkeland@nhh.no.

1 Introduction

An equal division of monetary rewards is a frequent outcome in many laboratory experiments in bargaining, and it is a common principle in many real life situations, for example, bequests to children (Camerer, 2003; Wilhelm, 1996). In other situations, for example, in bargaining over the output from production, experiments in economics and psychology have shown that many people follow a principle of proportionality, although a minority still prefer the equal division principle (Konow, 1996; Wagstaff, 2001; Gächter and Riedl, 2005; Cappelen, Sørensen, and Tungodden, 2010). Fairness principles such as equality and proportionality may arise from moral or political philosophy or simply be accepted over time as a way of dealing with distributive issues.

Hirschman (1977) and Elster (1989) have pointed to the fact that bargaining between individuals who strongly believe in different fairness principles can easily lead to conflict. They both argue that material self-interest can moderate conflicts of fairness principles, in the words of Elster (1989):

> The last case, norm conflict, is less likely to yield negotiated solutions. In norm-free bargaining, the only thing at stake is self-interest, a mild if mean-spirited passion. In norm conflict, the parties argue in terms of their honour, a notoriously strong passion capable of inspiring self-destructive and self-sacrificial behaviour. ... Compromises are possible between opposing norms, if one or both parties pour some water in their wine and let self-interest override honour. (Elster, 1989, p. 244).

In this paper, we develop a model that captures the potential conflict between two individuals who follow different fairness principles in bargaining. An individual's preferences are represented by a utility function where he or she trades off material self-interest and deviations from a fairness principle. The model is a variation of the frequently used inequity aversion model, which assumes that bargainers agree on a principle of equal division (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Bruyn and Bolton, 2008). The model in this paper builds on Cappelen, Hole, Sørensen, and Tungodden (2007), which allows individuals to follow different fairness principles. This introduces a dimension of conflict between two bargainers who follow different fairness principles, in addition to the trade-off between material self-interest and fairness motivation.[1]

[1] It is acknowledged in many models that different fairness principles should be considered, but are left out for reasons of intractability. See, for example, Andreoni and Bernheim (2009) who discuss fairness norms other than the 50–50 norm: 'If the players are asymmetric with respect to publicly observed inertia of merit, the fairness of an outcome might depend on the extent to which it departs from some other benchmark, such as $x^F = 0.4$. Provided the players agree on x^F, similar results would follow, except that the behavioural norm would correspond to the alternate benchmark. However, if players have different views of x^F, matters are more complex' (footnote 12).

Fairness motivation can influence both the possibility of reaching an agreement in bargaining and it can influence the properties of the agreement that is reached. The first part of the paper studies how fairness motivation influences the possibility of reaching an agreement. Proposition 2 formalizes the intuition of Elster (1989), that bargaining between two individuals who strongly believe in different fairness principles ends in conflict. This result shows the importance of considering a plurality of fairness principles to understand many bargaining problems. In contrast, Proposition 3 formalizes that if two bargainers follow the same fairness principle, it is always possible to reach an agreement.

The second part of the paper analyses the properties of the agreements that can be reached. We apply the Nash bargaining solution to bargaining situations between different types of individuals. We find that bargaining between an individual with strong fairness motivation and an individual motivated only by material self-interest, reaches an agreement that is closer to the fairness motivated individual's principle. If two bargainers who are motivated by fairness, but who disagree about what represents a fair division, reach an agreement, it will be a compromise between the two fair shares. We also show that in a bargaining situation where both individuals follow a fairness principle of strict equality, the Nash bargaining solution gives an equal division, and the trade-off between material self-interest and fairness motivation does not influence the solution.

Empirical studies have found that people follow a plurality of fairness principles in negotiations, and that this can explain bargaining impasses and how these are solved (Bazerman, 1985; Babcock, Loewenstein, Issacharoff, and Camerer, 1995; Babcock, Wang, and Loewenstein, 1996). Section 2 discusses data from a bargaining experiment that shows that it is also important to include a plurality of fairness principles to understand the properties of the agreements that are reached in bargaining. The agreements are from a bargaining experiment where participants have individually produced the endowment before they bargain over a division of the endowment. The two most common models for bargaining problems, material self-interest and preferences for equality, do not explain the experimental data in Section 2.

The model is presented in Section 3. Bargaining is then analysed in two steps in Section 4. First, the influence of fairness on the bargaining set is discussed without relying on a specific solution concept, and second, the Nash bargaining solution is applied to the problem. Section 5 contains some concluding remarks.

2 Experiment

Data from a laboratory experiment illustrate the importance of including fairness principles other than equality in bargaining.[2] The experiment consists of a

[2] This experiment is discussed in more detail in (Birkeland, 2011).

production phase and a bargaining phase. First, participants produce individually an output by typing a text from a transcript on the computer, and they receive a monetary reward equal to each correct word typed, rounded off to the nearest 50 words, multiplied by a randomly assigned high or low price. Second, participants were randomly matched into pairs and instructed to bargain over the endowment, which in this experiment is the sum of the individual production values. The experiment used an alternating offer bargaining protocol with infinite horizon.[3]

The 112 bargaining outcomes are shown in Figure 1, where the share of the total production value to person one and to person two are on the axes. The left panel shows the outcomes from 15 situations where both bargainers have produced the same amount and have the same price, and the right panel shows the outcomes from 97 situations where there is a difference in either the amount produced or the price. All the points that are along the diagonal from the lower left corner to the upper right corner are equal splits of the production value.

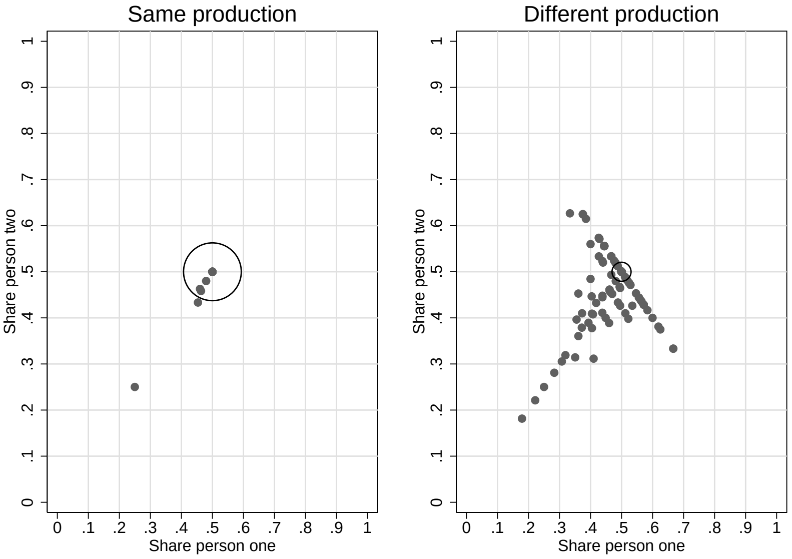
The results show that in all of the situations where bargainers have produced the same amount (left panel), the bargaining outcome is an equal division (the circle indicates the 10 observations that are equal divisions of the initial production value).[4] This result is consistent with the standard solution for players motivated by material self-interest. Alternating offer bargaining between players who are motivated by material self-interest, and have equal discount factors, gives an almost equal split (Rubinstein, 1982).

In the right panel, 49% of the outcomes are equal divisions that give the same amount of money to both participants. The number of equal divisions is significantly reduced when there are differences between the players from the production phase of the experiment ($p < 0.001$, Wilcoxon rank-sum test). These observations are inconsistent with the prediction of the standard model where bargainers are only motivated by material self-interest. We show later in the paper that a model where bargainers are motivated by strict equality cannot explain these observations either. A likely explanation for this shift to a more unequal division of the production value is that bargainers are motivated by fairness principles that do not imply equal division in these situations. A post-experimental questionnaire confirmed this hypothesis: 96% of the participants

[3] The alternating offer protocol starts with one of the players being randomly assigned as the first mover who suggests an opening offer in the first round ($t = 1$). Individual i proposes an amount of pay-off x_i for himself and $Y - x_i$ for the other player in each round of bargaining. The second mover responds to the opening offer by either accepting it and the bargaining is closed without cost, or by giving a counter offer in a second round ($t = 2$). The endowment shrinks in each round t by a discount factor δ_i^t. An agreement is reached when one player accepts the offer from the other player. In the experiment discussed in this paper, both players were induced with an equal discount factor, $\delta = 0.96$, which is so high that there is an insignificant first-mover advantage.

[4] To accommodate rounding to the nearest NOK 5, all agreements within the 47.5–52.5 split range are characterized as equal splits.

Figure 1: Experimental bargaining results



Note: The left panel shows outcomes from 15 situations where bargainers have the same production value, and the right panel shows bargaining outcomes from 97 situations where there are differences between bargainers in terms of either the amount produced or the price. The circle indicates the number of observations that are exactly a 50–50 split of the initial production value.

supported principles that justify unequal division in these situations, whereas only 4% of the participants supported equal division. Thus, the experiment emphasizes the importance of allowing for fairness principles other than equal division in economic models of bargaining to understand better many bargaining problems.

3 Model

In this section, we describe the theoretical framework for the analysis, including the bargaining environment and a utility function that can accommodate bargainers who are motivated by different fairness principles. We consider a bargaining environment in which two players bargain over how to divide an endowment, Y. Players can agree on any pair $x = (x_1, x_2)$ of shares of the endowment, $x_i \in [0, 1]$, such that the pair of shares is in the set $X = \{x \mid x_1 + x_2 \leq 1\}$, which is called the set of feasible agreements. In the following, we assume complete information, that is, the rules of the game and the utility functions of both players are common knowledge.

3.1 Fairness principle

An individual is assumed to have preferences that can be represented by a utility function where deviation from a fair share of the endowment reduces utility. The fair share of the endowment to individual i, according to his fairness principle k, is denoted as $s^{k(i)} \in [0,1]$. We assume that the fairness principle gives a unique division of the endowment, which is the case for all the fairness principles discussed in this paper.[5]

Assumption 1. *For any endowment, Y, and fairness principle, k, there exists a unique fair division $(s^{k(1)}, s^{k(2)})$, such that $s^{k(1)} + s^{k(2)} = 1$.*

The following example illustrates how principles of fairness could be applied in a production context. Consider a case where the endowment, Y, is the sum of individual production values, y_i, which can be decomposed into the individual production of units, e_i, and price, p_i, such that $y_i = e_i p_i$. In a two-person case, let $e_1 = 3$, $e_2 = 1$, $p_1 = \frac{1}{3}$, and $p_2 = 3$, then the production values are $y_1 = 1$ and $y_2 = 3$, and the endowment $Y = 4$. This could, for example, be bargaining between two executives about their share of a bonus in a corporation where one business area is exposed to the oil price and another business area is exposed to the aluminium price. In this context, there are three different distributive principles that are salient, $k = E, L, P$. The first fairness principle is an equal sharing of the monetary rewards, *strict equality*, which implies a fair share $s^{E(i)} = \frac{1}{2}$. The second fairness principle is a *laissez-faire* principle, where the individual production values determine the fair share to individual i, $s^{L(i)} = \frac{e_i p_i}{Y}$. The third principle is *proportionality*, where the fair share to individual i is proportional to the level of the production of units, $s^{P(i)} = \frac{e_i}{e_1+e_2}$, but where prices have no influence on the division.

In our example, implementation of these three fairness principles for two individuals gives the nine combinations of fair shares in Table 1. Each entry in the table shows the fair share that person one and person two claim according to their fairness principle, if that is only what they care about. The combinations of fairness principles can be divided into three categories: (i) both players follow the *same* fairness principle (diagonal elements), where by Assumption 1 the fair shares are always compatible; (ii) players one and two follow different fairness principles such that fair shares are incompatible, $s^{k(1)} + s^{k(2)} > 1$; (iii) players one and two follow different fairness principles such that fair shares sum to less than one, $s^{k(1)} + s^{k(2)} < 1$. Category (ii) is a natural bargaining situation where there

[5] The relevant set of fairness principles must be specified for the context in the model is applied. Which fairness principle an individual follows in a particular context, may depend on his identity; for example, an individual may follow a different fairness principle if he is an employer or if he is an employee (Akerlof and Kranton, 2010). The formulation of a fairness principle as a fair share of the endowment excludes some possible fairness principles, for example, principles that are related to the size of the endowment.

Table 1: Combinations of fairness principles

	$s^{E(2)}$	$s^{L(2)}$	$s^{P(2)}$
$s^{E(1)}$	$(\frac{1}{2}, \frac{1}{2})$	$(\frac{1}{2}, \frac{3}{4})$	$(\frac{1}{2}, \frac{1}{4})$
$s^{L(1)}$	$(\frac{1}{4}, \frac{1}{2})$	$(\frac{1}{4}, \frac{3}{4})$	$(\frac{1}{4}, \frac{1}{4})$
$s^{P(1)}$	$(\frac{3}{4}, \frac{1}{2})$	$(\frac{3}{4}, \frac{3}{4})$	$(\frac{3}{4}, \frac{1}{4})$

Note: The table shows combinations of the fairness principles E, L, P for person one and person two for parameters $e_1 = 3, e_2 = 1, p_1 = \frac{1}{3}, p_2 = 3$. Each entry shows the fair shares that person one and person two claim according to their principles if they only care about fairness.

is conflict of interest. Category (iii) is less important in bargaining and will not be discussed in the following analysis.

In this numerical example, different fairness principles give different fair shares, but this is not necessarily the case in all situations. Different principles can also give the same fair shares in some situations; for example, an equal division follows both from the fairness principle of strict equality, and from the principle of proportionality if both individuals have produced the same number of units.

3.2 Utility function

We assume that the utility function is additively separable for individual i in his own share of the endowment, x_i, and the cost of deviating from the fair share, $x_i - s^{k(i)}$. The endowment is assumed to be non-negative, $Y \geq 0$.

Assumption 2. *Individual i's preferences can be represented by the utility function:*

$$u_i(x_iY, s^{k(i)}Y) = (x_i - \beta_i(x_i - s^{k(i)})^2)Y.$$

The utility loss from deviating from the fair share is squared, which implies that the utility loss from deviation to the better or the worse is symmetric, and that the utility loss increases exponentially with the distance. The weight individual i has on not deviating from his fair share is given by the parameter β_i, which is assumed to be non-negative, $\beta_i \geq 0$. If $\beta_i = 0$, the model is reduced to material self-interest, a utility function that is linear in x_i. The parameter β_i captures tension between an individual's motivation to follow a fairness principle and that of material self-interest.[6] This utility function is continuous, twice

[6] The functional form of the utility function follows the two-person case of Bolton and Ockenfels (2000), Lopomo and Ok (2001), Cappelen et al. (2007), and Bruyn and Bolton (2008). The parameter β_i captures the trade-off that is represented by the fraction $\frac{b_i}{2a_i}$ in Bolton and Ockenfels (2000). The utility function in this paper differs from Bolton and Ockenfels (2000)

differentiable and concave in x_i. The utility function attains its inner maximum when:

$$x_i^* = \frac{1}{2\beta_i} + s^{k(i)}.$$

The interior solution to an individual, x_i^*, is not defined for $\beta_i = 0$, and it is independent of the endowment Y. The utility function is strictly increasing in the interval $0 \leq x_i \leq s^{k(i)}$. Only for small values of β_i, that is, $\beta_i < \frac{1}{2(1-s^{k(i)})}$, is the utility function strictly increasing in the entire interval $0 \leq x_i \leq 1$. For high values of β_i, the importance of obtaining the fair share outweighs the utility of obtaining a larger share of the endowment, see also Bolton and Ockenfels (2000). In the case that $\beta_i \to \infty$, the interior solution approaches the fair share, $x_i^* \to s^{k(i)}$.

This model allows for different types of players: a material self-interested player, a strongly fairness motivated player, as well as an intermediate type that trades off material self-interest and fairness motivation. It also introduces differences regarding which fairness principle players follow.

The effects of changing the parameters of the weights assigned to fairness, β_i, and the fair share, $s^{k(i)}$, are illustrated in Figure 2. We observe that all three utility functions illustrated in the figure attain negative values for a small x, but only the utility function with a high β_i attains negative values for a large x. Shifting the fair share, $s^{k(i)}$, to a higher value raises the point where the utility function is zero. We also observe that the utility function with a high β_i envelops the utility function with the same fair share, $s^{k(i)}$, but a lower β_i, because an increase in fairness motivation reduces the utility from deviating from the fair share.

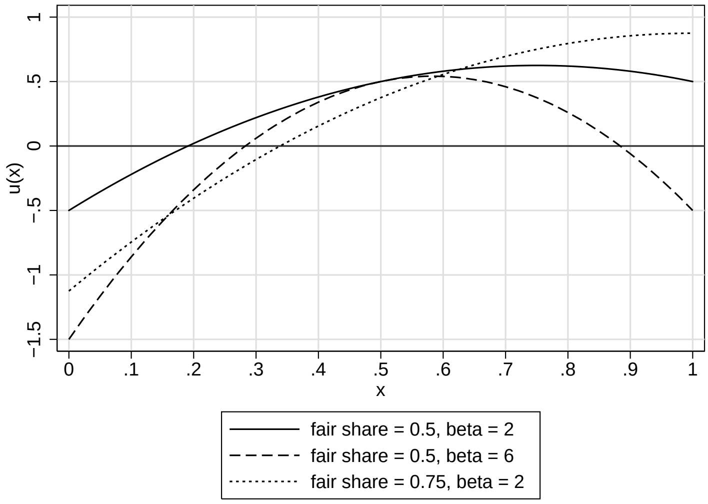
In bargaining, people evaluate their utility from possible agreements against the utility from the situation where no agreement is reached. We assume that the disagreement utility is zero.

Assumption 3. *Individual i's utility from disagreement is zero, $u_i^d = 0$.*

Assumption 3 follows directly from the specification of the utility function in an economic environment where the endowment is zero in disagreement, $Y = 0$, which is the case in many bargaining experiments. This assumption implies,

in that it allows for fair shares other than $s^{k(i)} = \frac{1}{2}$. In the model of Lopomo and Ok (2001), a bargainer's utility depends on both his absolute gain and the relative share he gets compared with the average share. The model of Lopomo and Ok (2001) allows for uncertainty about the weight that the other players have on the deviation from the average share. The utility function in Cappelen et al. (2007) allows for different fairness principles, but the principles are not defined in shares of the endowment. The utility function used in Bruyn and Bolton (2008) is linear above an equal division. If you introduce players with different fairness principles in a model with asymmetries in valuing deviations from a fair share, e.g., the Bruyn and Bolton (2008) model, this could easily result in a non-convex Pareto frontier of the bargaining set, which could give multiple Nash bargaining solutions.

Figure 2: The utility function



Note: The utility function given in Assumption 2 for different parameter values, $Y = 1$.

however, that an individual's fairness consideration does not apply to the disagreement outcome.

3.3 Reservation points

An individual's reservation point is defined as the share that makes an individual indifferent between accepting an offer or choosing disagreement, which gives zero utility. As illustrated in Figure 2, the present model allows for more than one reservation point. The lower reservation point, x_i^L in the interval $0 \leq x_i^L \leq s^{k(i)}$, where $u_i(x_iY, s^{k(i)}Y) = 0$ is given by:

$$x_i^L = \frac{1 + 2\beta_i s^{k(i)} - \sqrt{1 + 4\beta_i s^{k(i)}}}{2\beta_i}.$$

The reservation point is influenced both by the fairness principle and the weight that an individual attaches to following the fairness principle. For a fairness motivated individual, $\beta_i > 0$, an offer below the lower reservation point would be considered too unfair, and it is therefore rejected, although the offer represents a positive share of the endowment. In contrast, the utility function for an individual who is only motivated by material self-interest, $\beta_i = 0$, always has a lower reservation point at zero, and he would not reject a positive share of the endowment.

An important property of the lower reservation point is that when β_i increases the reservation point, x_i^L, approaches the fair share:

$$\lim_{\beta_i \to \infty} x_i^L = s^{k(i)}.$$

We focus on the lower reservation because it is used later in the paper to analyse the bargaining set. For large values of β_i, there is also an upper reservation point, x_i^H, in the interval $s^{k(i)} \leq x_i^H \leq 1$. The upper reservation point will correspond to situations where a player, for example, is offered the whole gain, and he rejects this as unfair even though it benefits him.

Proposition 1. Reservation points. *For any fair share $s^{k(i)}$, there is a value $\hat{\beta}_i$ such that for any $0 < \beta_i < \hat{\beta}_i$ there exists a unique reservation point, x_i^L, and for any $\beta_i \geq \hat{\beta}_i$ there exist two reservation points, x_i^L, and x_i^H.*

Proof. See Appendix A. □

4 Bargaining solutions

In this section, we use cooperative bargaining theory to analyse the outcome from bilateral bargaining with fairness motivated individuals. A cooperative bargaining solution is not based on a specific bargaining process. It is assumed that an agreement from bargaining is binding and enforceable through a legal system outside of the model. Cooperative bargaining theory builds on a utility representation of the feasible agreements, which is the convex utility set:[7]

$$\mathcal{U} = \left\{ (u_1(x_1 Y, s^{k(1)} Y), u_2(x_2 Y, s^{k(2)} Y)) : x \in X \right\}.$$

First, following standard analysis we require that the agreement is at the frontier of the utility set, which implies that the agreement is in the set $Z = \{x \mid x_1 + x_2 = 1\}$.[8] Second, we require that the agreement gives both players at least as much utility as they can get without an agreement, which from Assumption 3 implies that the agreement must give both players at least zero utility. These two requirements ensure that the agreement is in the *bargaining set*, $\mathcal{B}$.[9]

Assumption 4. *An agreement is in the bargaining set:*

$$\mathcal{B}(\mathcal{U}) = \left\{ (u_1(x_1 Y, s^{k(1)} Y), u_2(x_2 Y, s^{k(2)} Y)) \geq (0,0) : x \in Z \right\}.$$

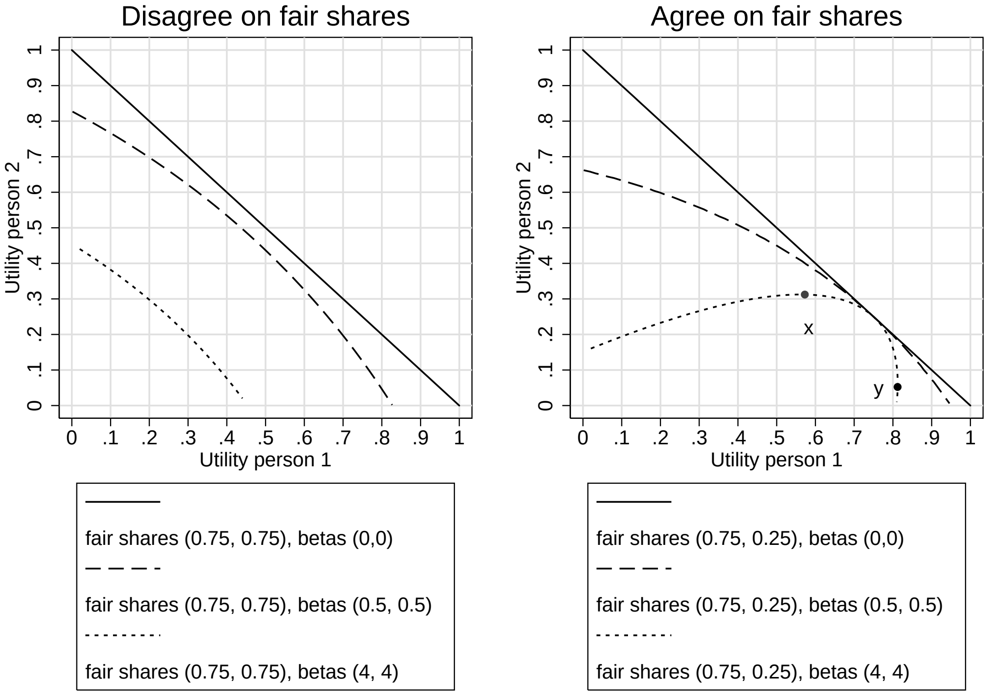
Bargaining sets for two individuals are illustrated in Figure 3, where the left panel shows individuals who disagree on what is a fair division, and the right panel shows individuals who agree on what is a fair division. Each line represents a bargaining set for different parameter values for person one and person two. The two end-points of a bargaining set are defined where person one gets maximum utility given that person two gets enough utility to accept the agreement, which must be at his lower reservation point, $(u_1^{max}, u_2(x_2^L) = 0)$, and similarly where person two gets maximum utility given that person one gets enough utility to accept the agreement $(u_1(x_1^L) = 0, u_2^{max})$.

[7] Two strictly increasing concave functions give a convex combination (Binmore, 2007).

[8] Roth (1979) discusses the standard assumptions of cooperative bargaining theory.

[9] Here, $\geq$ is defined coordinatewise, that is, $(x_1, x_2) \geq (y_1, y_2)$ iff $x_i \geq y_i$ for each $i = 1, 2$.

Figure 3: Bargaining sets



Note: The left panel shows bargaining sets where players have incompatible fair shares, and the right panel shows bargaining sets where they agree on fair shares. The lines represent different levels of (β_1, β_2). The endowment $Y = 1$. Points marked x and y are the end-points of the Pareto frontier for the dotted line.

We observe that the bargaining set for fairness motivated individuals is typically smaller than the bargaining set for players who are more motivated by material self-interest. The bargaining set shrinks when β_i increases because an individual gets disutility from deviating from the fair share, and he consequently increases his reservation point. A smaller bargaining set means that there are fewer possible agreements that can be realized. It is often argued that the frequency of disagreements in bargaining is higher when there are fewer possible agreements that can be realized. The relationship between the size of the bargaining set and the efficiency of bargaining is, however, unclear. Crawford (1982), for example, develops a bargaining model where disagreements are reduced when the size of the bargaining set shrinks.

The left panel in Figure 3 represents bargaining sets where players have incompatible fair shares. In this case, the bargaining set moves towards the origin for bargainers who are more fairness motivated, and at some point, the bargaining set is empty, which occurs when reservation points are incompatible, i.e., the reservation points combined constitute more than the endowment. Consequently, players prefer to disagree. This shows that strong preferences for conflicting fairness principles make it impossible to reach an agreement.

Proposition 2. Principled disagreement. *If the fair shares are incompatible,* $s^{k(1)} + s^{k(2)} > 1$*, then there exists a* $(\hat{\beta}_1, \hat{\beta}_2)$ *such that for any* $(\beta_1 \geq \hat{\beta}_1, \beta_2 \geq \hat{\beta}_2)$*, the only feasible solution is the disagreement outcome.*

Proof. See Appendix A. □

The right panel in Figure 3 represents bargaining sets where both players follow the same fairness principle. We observe from Figure 3 that for players who are more fairness motivated, the bargaining set shrinks and envelopes the point that both players consider a fair division. The bargaining set shrinks further for players who are strongly motivated by the fairness principle, and in the limit, the bargaining set only contains the fair solution. Proposition 3 formalizes the point that, for bargainers who follow the same fairness principle, it is always possible to reach an agreement.

Proposition 3. Principled agreement. *If two individuals follow the same fairness principle,* k*, then there always exists a non-empty bargaining set,* $\mathcal{B}$*. Increases in* β_1 *and* β_2 *give a shrinking bargaining set, and in the limit, it collapses to a single point, which represents the fair division* $(s^{k(1)}, s^{k(2)})$*.*

Proof. See Appendix A. □

The reservation point of a player can drop below the other player's utility maximizing offer, and still give the first player more utility than in disagreement. This occurs when the two bargainers agree on the fairness principle, and at least one of them is strongly fairness motivated. The bargaining set marked with a dotted line in the right panel in Figure 3 has two points marked x and y. The line connecting these two points is the Pareto frontier.[10] A high β_i changes the curvature of the frontier of the bargaining set such that the line segment up to the point marked x represents a Pareto improvement for player one, and the line segment up to the point marked y represents a Pareto improvement for player two. Pareto optimality is a requirement for the Nash bargaining solution discussed in the next section.

[10] The point marked x in Figure 3 is defined as the maximum utility that player two can achieve, given that no further Pareto improvement for player one is possible, $(\tilde{u}_1, u_2^{max})$ where $\tilde{u}_1 > u_1(x^L) = 0$, and the point marked y is the maximum utility that player one can achieve, given that no further Pareto improvement for player two is possible, $(u_1^{max}, \tilde{u}_2)$, where $\tilde{u}_2 > u_2(x^L) = 0$.

4.1 Nash bargaining solution

A commonly used concept for finding a unique outcome in the bargaining set is the Nash bargaining solution (Nash, 1950).[11] The Nash bargaining solution is the maximum of the product of the utility minus the utility of disagreement, u_i^d:

$$max(u_1 - u_1^d)(u_2 - u_2^d).$$

The analytical solution to the Nash bargaining solution for the model developed in this paper is derived in Appendix B. The non-linearity of the utility function makes the analytical solution difficult to interpret, but the effect of changing the parameters can easily be interpreted by studying numerical computations.

Figure 4 shows the Nash bargaining solution for different combinations of players. Each point on the four panels shows the share that player one receives, x_1, at different levels of the fairness weight, β_1.[12] The four panels show matching of player one against different types of player two. The standard solution for two players only motivated by material self-interest is an equal division of the monetary gain, which is the starting point in the upper left panel. This panel shows that if player one has a higher weight on following his fairness principle, the bargaining solution gives a share that is closer to his fair share, which in this example is $s^{k(1)} = \frac{3}{4}$. A fairness motivated player who takes a principled stand in bargaining will achieve a solution that is closer to his fairness principle if he bargains against a player who is only motivated by material self-interest.

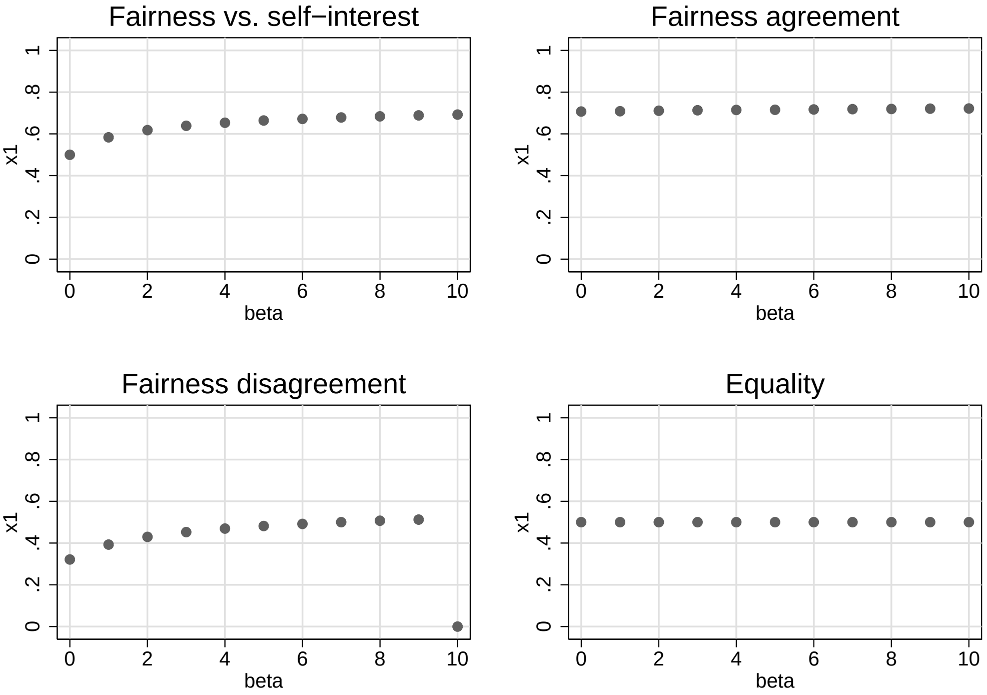
In the upper right panel both players are fairness motivated, and they agree on the fair division ($s^{k(1)} = \frac{3}{4}, s^{k(2)} = \frac{1}{4}$). We observe that all the bargaining solutions are close to the fair division. An increase in the trade-off between self-interest and fairness motivation, β_1, has an insignificant effect on the bargaining solution. In line with Proposition 2, sufficiently high weights on following the same fairness principle give the fair division.

In the lower left panel, player two is also a fairness motivated player, but in this example player two disagrees with player one about the fairness principle. Both players believe that it would be fair if they get three-quarters of the endowment; thus, they both have the same fair share ($s^{k(1)} = \frac{3}{4}, s^{k(2)} = \frac{3}{4}$). From the lower left panel, we see that at low levels of β_1, the bargaining solution is a division that is close to player two's fair share. At high levels of β_1, the bargaining solution is a compromise solution between the fair shares, which in this example is an equal

[11] Nash (1950) proves that this is the only solution that fulfils four reasonable axioms: (i) the solution should be independent of affine transformations of the utility function; (ii) the solution should be independent of irrelevant alternatives; (iii) the solution should treat players symmetrically; and (iv) the solution should be Pareto optimal. An introduction to bargaining theory and the Nash axioms is found in Roth (1979) and Binmore (2007).

[12] Bruyn and Bolton (2008) and Cappelen et al. (2007) estimate the average weight that an individual has on following his fairness principle. The average weight, converted to a value comparable to β_i, is 6.0 for a three-round bargaining game in Bruyn and Bolton (2008), and 7.7 for a dictator game in Cappelen et al. (2007).

Figure 4: Nash bargaining solution



Note: The graph shows the Nash bargaining solution for the share that player one receives, x_1, for different levels of the fairness weight, β_1. Player one has a fairness principle $s^{k(1)} = \frac{3}{4}$ in the first three panels, and $s^{k(1)} = \frac{1}{2}$ in the lower right panel. The parameters for player two are: 'Fairness vs. self-interest': $\beta_2 = 0$; 'Fairness agreement': $\beta_2 = 7$, $s^{k(2)} = \frac{1}{4}$; 'Fairness disagreement': $\beta_2 = 7$, $s^{k(2)} = \frac{3}{4}$; 'Equality': $\beta_2 = 7$, $s^{k(2)} = \frac{1}{2}$. The endowment $Y = 1$.

division. Importantly, in line with Proposition 3, two players who disagree about what is a fair share cannot reach an agreement if their fairness motivation is too strong. At the level of $\beta_1 = 10$ in the lower left panel, the bargain solution is the disagreement outcome of zero.

The lower right panel shows the case where both players are fairness motivated and both players agree on a principle of strict equality. We can see that, in this case, the trade-off between self-interest and fairness motivation, β_1, does not influence the solution. This last result follows from the property of symmetric treatment of players in the Nash bargaining solution.

Proposition 4. Fairness weight impotency. *If bargainers follow the fairness principle of strict equality, $s^{k(1)} = s^{k(2)} = \frac{1}{2}$, then the Nash bargaining solution is $(x_1^N = \frac{1}{2}, x_2^N = \frac{1}{2})$ for any β_1, β_2.*

Proof. See Appendix A. □

4.2 The generalized Nash bargaining solution

There is a version of the Nash bargaining solution that allows for bargaining power, α_i, to influence the solution:

$$max(u_1 - u_1^d)^{\alpha_1}(u_2 - u_2^d)^{\alpha_2}.$$

In the standard case where individuals are only motivated by material self-interest, the individual with more bargaining power gets a larger share of the endowment than a player with less bargaining power. Similarly, in bargaining between two equally fairness motivated individuals who disagree about fair shares, the individual with more bargaining power gets closer to his fair share. This argument also works the other way, a more fairness motivated individual gets closer to his fair share for a given distribution of bargaining power. Fairness motivation can therefore counterbalance the influence of unfavourable bargaining power.

Moreover, the same outcome that follows from bargaining between two equally strong fairness motivated individuals who agree on the fairness principle, may also be the result of bargaining between self-interested individuals who have bargaining power distributed in the same proportion as the fair shares. An interpretation of this result is that there are two ways to achieve a fair outcome in bargaining, through agreement about fairness principles or regulation of bargaining power.

However, in one case where the relative bargaining power is distributed in the exact same proportion as the fair shares that follow from a fairness principle, fairness motivation does not influence the generalized Nash bargaining solution. This result is similar to Proposition 4 where individuals are motivated by the fairness principle of strict equality and they have equal bargaining power. Numerical computation gives support to the following conjecture (see Appendix A).

Conjecture 1. Generalized fairness weight impotency. *If $s^{k(1)} + s^{k(2)} = 1$, $\alpha_1 = s^{k(1)}$ and $\alpha_2 = s^{k(2)}$, then the generalized Nash bargaining solution is $(x_1^N = s^{k(1)}, x_2^N = s^{k(2)})$, for any β_1, β_2.*

5 Concluding remarks

Individuals who are only motivated by material self-interest are always able to make a compromise and find an agreement, provided that the monetary reward from agreement is higher than from the disagreement outcome. For fairness motivated individuals, the outcome from bargaining will depend both on the principle they follow, and the trade-off they make between following the principle and material self-interest. First, people who are motivated by the same fairness principle have a non-empty bargaining set and it is always possible to reach an agreement, and if they have a high weight on following the principle, they will agree on the fair outcome. Second, if people disagree about what is a fair share, it may be impossible to reach an agreement, particularly if they have a high weight on following their principles. Disagreement can easily be the outcome from bargaining between players that insist on different fairness principles.

The Nash bargaining solution shows that bargaining between an individual with strong fairness motivation and an individual only motivated by material self-interest reaches an outcome that is closer to the fairness motivated individual's principle. In bargaining between two individuals motivated by fairness, and who disagree about what represents a fair division, the Nash bargaining solution gives an outcome that is a compromise between the fair shares. If individuals follow the commonly assumed fairness principle of strict equality, the trade-off between following the fairness principle and material self-interest does not influence the outcome. The generalized Nash bargaining solution shows that a strongly fairness motivated individual can balance the higher bargaining power of another individual.

This research could be extended both theoretically and empirically. An interesting theoretical extension is to incorporate characteristics of the disagreement outcome into individuals' fairness principles. Another important issue for further research is how social preference models influence the efficiency of bargaining.

Finally, I would like to point to several empirical hypotheses for fairness motivation that can be derived from the analysis in this paper. First, material self-interest may be more predominant in societies where there is a great deal of plurality of fairness principles among people, and conversely, in a more homogeneous society, people may be more fairness motivated. In societies with a great deal of heterogeneity, it is important to reach compromises in transactions. Thus, an environment that fosters material self-interest may perform better than one that fosters fairness motivation. Second, the analysis in this paper shows that fairness motivation could develop among groups in societies where the bargaining power is to their disadvantage. The mobilization of fairness motivation can neutralize the imbalance of bargaining power. The development of strong fairness principles among unions in wage negotiations could be an example of this. Third, in societies with strong groups that are motivated by different fairness principles, the analysis shows that there can be more conflicts, for example, in societies where employers and employees strongly believe in different principles of wage setting.

References

Akerlof, George A. and Rachel E. Kranton (2010). *Identity Economics*, Princeton University Press.

Andreoni, James and Douglas Bernheim (2009). "Social image and the 50-50 norm: A theoretical and experiemental analysis of audience effects", *Econometrica*, 77: 1607–1636.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer (1995). "Biased judgements of fairness in bargaining", *American Economic Review*, 85(5): 1337–1343.

Babcock, Linda, Xianghong Wang, and Georg Loewenstein (1996). "Choosing the wrong pond: Social comparisons in negotiations that reflect a self-serving bias", *The Quarterly Journal of Economics*, CXI(1): 1–19.

Bazerman, Max H. (1985). "Norms of distributive justice in interest arbitration", *Industrial and Labor Relations Review*, 38(4): 558–570.

Binmore, Ken (2007). *Playing for Real*, Oxford University Press.

Birkeland, Sigbjørn (2011). "Negotiation under possible third party settlement", Norwegian School of Economis and Business Administration, discussion paper.

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

Bruyn, Arnaud De and Gary E. Bolton (2008). "Estimating the influence of fairness on bargaining behavior", *Management Science*, 54(10): 1774–1791.

Camerer, Colin (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press.

Cappelen, Alexander W., Astrid D. Hole, Erik Ø. Sørensen, and Bertil Tungodden (2007). "The pluralism of fairness ideals: An experimental approach", *American Economic Review*, 97(3): 818–827.

Cappelen, Alexander W., Erik Ø. Sørensen, and Bertil Tungodden (2010). "Responsibility for what? fairness and individual responsibility", *European Economic Review*, 54(3): 429–441.

Crawford, Vincent P. (1982). "A theory of disagreement in bargaining", *Econometrica*, 50: 607–38.

Elster, Jon (1989). *The Cement of Society*, Cambridge University Press.

Fehr, Ernst and Klaus M. Schmidt (1999). "A theory of fairness, competition and cooperation", *The Quarterly Journal of Economics*, 114(3): 917–868.

Gächter, Simon and Arno Riedl (2005). "Moral property rights in bargaining with infeasible claims", *Management Science*, 51(2): 249–263.

Hirschman, Albert O. (1977). *The Passions and the Interests : Political Arguments for Capitalism before Its Triumph*, Princeton University Press.

Konow, James (1996). “A positive theory of economic fairness”, *Journal of Economic Behaviour and Organisation*, 31: 13–35.

Lopomo, Giuseppe and Efe A. Ok (2001). “Bargaining, interdependence, and the rationality of fair division”, *RAND Journal of Economics*, 32: 263–283.

Nash, John F. (1950). “The bargaining problem”, *Econometrica*, 18: 155–62.

Roth, Alvin E. (1979). *Axiomatic Models of Bargaining*, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag.

Rubinstein, Ariel (1982). “Perfect equilibrium in a bargaining model”, *Econometrica*, 50: 97–109.

Wagstaff, Graham F. (2001). *An Integrated Psychological and Philosophical Approach to Justice*, The Edwin Mellen Press.

Wilhelm, Mark O. (1996). “Bequest behavior and the effect of heirs’ earnings: Testing the altruistic model of bequests”, *The American Economic Review*, 86: 874–892.

Appendix A Proof of propositions

Proof of Proposition 1

We want to show that for any fair share, $s^{k(i)}$, there is a value $\hat{\beta}_i$ such that for any $0 < \beta_i < \hat{\beta}_i$ there exists a unique reservation point, x_i^L, and for any $\beta_i \geq \hat{\beta}_i$ there exist two reservation points, x_i^L, and x_i^H. The utility function in Assumption 1:

$$u_i(x_iY, s^{k(i)}Y) = (x_i - \beta_i(x_i - s^{k(i)})^2)Y,$$

is a quadratic equation:

$$a_1x^2 + a_2x + a_3 = 0,$$

where the coefficients are reduced to:

$$\begin{aligned} a_1 &= -\beta_i, \\ a_2 &= (1 + 2\beta_i s^{k(i)}), \\ a_3 &= -\beta_i(s^{k(i)})^2. \end{aligned}$$

The discriminant, $a_2^2 - 4a_1a_3 = 1 + 4\beta_i s^{k(i)}$, is positive and hence the utility function has two real, distinct roots. The quadratic formula gives the two solutions:

$$x_i^L = \frac{1 + 2\beta_i s^{k(i)} - \sqrt{1 + 4\beta_i s^{k(i)}}}{2\beta_i}, \qquad x_i^H = \frac{1 + 2\beta_i s^{k(i)} + \sqrt{1 + 4\beta_i s^{k(i)}}}{2\beta_i}.$$

These solutions are not defined for $\beta_i = 0$. By definition, a fair share, $s^{k(i)}$, can have values in the interval $0 \leq s^{k(i)} \leq 1$. We see that if $s^{k(i)} = 0$, then $x_i^L = 0$. Differentiate x_i^L with respect to β_i:

$$\frac{dx_i^L}{d\beta_i} = \frac{1 + 2\beta_i s^{k(i)} - \sqrt{1 + 4\beta_i s^{k(i)}}}{2\beta_i^2 \sqrt{1 + 4\beta_i s^{k(i)}}}.$$

For the numerator to be positive, $1 + 2\beta_i s^{k(i)} > \sqrt{1 + 4\beta_i s^{k(i)}}$. By squaring both sides of the inequality we see that the numerator is always positive for $s^{k(i)} > 0$. Hence, $\frac{dx_i^L}{d\beta_i} > 0$, and x_i^L is strictly increasing in β_i for $s^{k(i)} > 0$. We know from Section 3.3 that $\lim_{\beta_i \to \infty} x_i^L = s^{k(i)}$. Thus, there always exists a lower reservation point, x_i^L, which attains values in the interval $0 \leq x_i^L \leq s^{k(i)}$.

We then consider the upper reservation point, x_i^H. We can see that if $s^{k(i)} = 1$, then $x_i^H > 1$, which is outside of the domain of the utility function for argument x_i. Differentiate x_i^H with respect to β_i:

$$\frac{dx_i^H}{d\beta_i} = \frac{-1 - 2\beta_i s^{k(i)} - \sqrt{1 + 4\beta_i s^{k(i)}}}{2\beta_i^2 \sqrt{1 + 4\beta_i s^{k(i)}}}.$$

We see that x_i^H is strictly decreasing in β_i, since $\frac{dx_i^H}{d\beta_i} < 0$. If $s^{k(i)} < 1$, then x_i^H may attain values in the interval $0 < x_i^H \leq 1$, depending on the relationship between $s^{k(i)}$ and β_i. Define $\hat{\beta}_i$ such that $x^H = 1$, which gives:

$$\hat{\beta}_i = \frac{1}{(1 - s^{k(i)})^2}.$$

Any $\beta_i \geq \hat{\beta}_i$ will give an upper reservation point in the interval $0 < x_i^H \leq 1$. Hence, for β_i in the interval $0 < \beta_i < \hat{\beta}_i$, there exists a unique reservation point, x_i^L, and for $\beta_i \geq \hat{\beta}_i$ there exist two reservation points, x_i^L, and x_i^H.

Proof of Proposition 2

We want to show that if the fair shares are incompatible, $s^{k(1)} + s^{k(2)} > 1$, then there exists a $(\hat{\beta}_1, \hat{\beta}_2)$ such that for any $(\beta_1 \geq \hat{\beta}_1, \beta_2 \geq \hat{\beta}_2)$, the only feasible solution is the disagreement outcome. Note that by assumption, if $s^{k(1)} + s^{k(2)} > 1$, then $s^{k(1)} > 0$ and $s^{k(2)} > 0$.

1. From Section 4 we know that $\mathcal{B}$ is empty if $x_1^L + x_2^L > 1$.

2. From Proposition 1 we know that there exists a lower reservation point, x_i^L, which is strictly increasing in β_i for $s^{k(i)} > 0$.

3. Define $\varepsilon < |\frac{1 - s^{k(1)} - s^{k(2)}}{2}|$. Find $(\hat{\beta}_1, \hat{\beta}_2)$ such that $x_1^L = s^{k(1)} - \varepsilon$ and $x_2^L = s^{k(2)} - \varepsilon$.

4. It follows from step 2 and step 3 that $x_1^L + x_2^L > 1$ for any $(\beta_1 \geq \hat{\beta}_1, \beta_2 \geq \hat{\beta}_2)$. Hence, from step 1 it then follows that $\mathcal{B}$ is empty, and the only feasible solution is the disagreement outcome.

Proof of Proposition 3

Consider any combination of $(\beta_1,\, \beta_2)$. We want to show that $\mathcal{B}(\beta_1,\, \beta_2)$ is non-empty, i.e. $x_1^L + x_2^L \leq 1$, if two individuals who follow the same fairness principle, k.

1. From Assumption 1 we know that if two individuals follow the same fairness principle, k, then $s^{k(1)} + s^{k(2)} = 1$.

2. From Proposition 1 we know that there exists a lower reservation point, x_i^L, which is monotonically increasing in β_i in the interval $0 \leq x_i^L \leq s^{k(i)}$. For a given $(\beta_1,\, \beta_2)$, we find (x_1^L, x_2^L) by using the formula in Proposition 1.

3. By taking into account step 1, it follows from step 2 that $x_1^L + x_2^L \leq 1$, and hence $\mathcal{B}(\beta_1,\, \beta_2)$ is non-empty. This completes the proof of the first part of Proposition 3.

4. We also want to show that an increase in β_1 and β_2 in the limit collapses to a single point, which represents the fair division $(s^{k(1)}, s^{k(2)})$. If $(\beta_1, \beta_2) \rightarrow \infty$, it follows from Proposition 1 that $(x_1^L, x_2^L) \rightarrow (s^{k(1)}, s^{k(2)})$. By Assumption 1, $(s^{k(1)}, s^{k(2)})$ is a unique element. Hence, in the limit $\mathcal{B}$ only contains one element, which is the utility representation of $(s^{k(1)}, s^{k(2)})$. This completes the proof of the second part of Proposition 3.

Proof of Proposition 4

Consider any combination of $\beta_1 \geq 0$ and $\beta_2 \geq 0$. We want to show that if bargainers follow the fairness principle of *strict equality*, then the Nash bargaining solution is $(x_1^N = \frac{1}{2}, x_2^N = \frac{1}{2})$.

1. Consider the case where $\beta_1 = \beta_2 = 0$. The model is then reduced to a standard utility function, and it follows straightforwardly that the solution is $(x_1^N = \frac{1}{2}, x_2^N = \frac{1}{2})$.

2. Consider the case where $\beta_1 > 0$ and $\beta_2 > 0$. By assumption, the parameter values are $s^{k(1)} = s^{k(2)} = \frac{1}{2}$ in the Nash bargaining solution.

3. Differentiate the Nash bargaining solution (as stated in Appendix B) with respect to β_1 and β_2. If you evaluate these two expressions for any β_1 and β_2, the outcome is zero. Hence, changes in β_1 and β_2 do not influence the Nash bargaining solution.

Examples of Conjecture 1

Table 2: Asymmetric bargaining power

			β_2	
		1	5	10
	1	(0.75, 0.25)	(0.75, 0.25)	(0.75, 0.25)
β_1	5	(0.75, 0.25)	(0.75, 0.25)	(0.75, 0.25)
	10	(0.75, 0.25)	(0.75, 0.25)	(0.75, 0.25)

Note: Table 2 shows a numerical computation of the generalized Nash bargaining solution for person one and person two for parameters $\alpha_1 = s^{k(1)} = 0.75$, and $\alpha_2 = s^{k(2)} = 0.25$.

Table 3: Symmetric bargaining power

			β_2	
		1	5	10
	1	(0.5, 0.5)	(0.5, 0.5)	(0.5, 0.5)
β_1	5	(0.5, 0.5)	(0.5, 0.5)	(0.5, 0.5)
	10	(0.5, 0.5)	(0.5, 0.5)	(0.5, 0.5)

Note: Table 3 shows a numerical computation of the generalized Nash bargaining solution for person one and person two for parameters $\alpha_1 = s^{k(1)} = 0.5$, and $\alpha_2 = s^{k(2)} = 0.5$.

Appendix B Analytical solution

The Nash bargaining solution can be found by solving the optimization problem:

$$\begin{aligned} \max \quad & (u_1 - u_1^d)(u_2 - u_2^d) \\ \text{s.t.} \quad & x_1 + x_2 = 1. \end{aligned}$$

From Section 3.2 we have:

$$\begin{aligned} & u_i(x_i Y, s^{k(i)} Y) = (x_i - \beta_i (x_i - s^{k(i)})^2) Y, \\ & u_i^d = 0. \end{aligned}$$

By substituting the constraint into the objective function, the optimization problem can be written as:

$$\max f(x_1) = \Big((x_1 - \beta_1 (x_1 - s_1)^2) Y \Big) \Big((1 - x_1 - \beta_2 (1 - x_1 - s_2)^2) Y \Big).$$

Differentiating f with respect to x_1 gives a cubic equation (the subscript on x is suppressed):

$$a_1x^3 + a_2x^2 + a_3x + a_4 = 0,$$

where the coefficients are:

$$\begin{aligned}
a_1 &= 4\beta_1\beta_2Y^2,\\
a_2 &= (3\beta_1 - 3\beta_2 - 6\beta_1\beta_2 - 6s_1\beta_1\beta_2 + 6s_2\beta_1\beta_2)Y^2,\\
a_3 &= (-2 - 2\beta_1 - 4s_1\beta_1 + 4\beta_2 - 4s_2\beta_2 + 2\beta_1\beta_2 + 8s_1\beta_1\beta_2 + 2s_1^2\beta_1\beta_2\\
&\quad - 4s_2\beta_1\beta_2 - 8s_1s_2\beta_1\beta_2 + 2s_2^2\beta_1\beta_2)Y^2,\\
a_4 &= (1 + 2s_1\beta_1 + s_1^2\beta_1 - \beta_2 + 2s_2\beta_2 - s_2^2\beta_2 - 2s_1\beta_1\beta_2 - 2s_1^2\beta_1\beta_2\\
&\quad + 4s_1s_2\beta_1\beta_2 + 2s_1^2s_2\beta_1\beta_2 - 2s_1s_2^2\beta_1\beta_2)Y^2.
\end{aligned}$$

Define the following relationships:

$$\begin{aligned}
Q &\equiv \frac{a_3}{3a_1} - \left(\frac{a_2}{3a_1}\right)^2,\\
R &\equiv \frac{a_3a_2}{6a_1^2} - \frac{a_4}{2a_1} - \left(\frac{a_2}{3a_1}\right)^3,\\
D &\equiv Q^3 + R^2,\\
S &\equiv \left(R + \sqrt{D}\right)^{\frac{1}{3}},\\
T &\equiv \left(R - \sqrt{D}\right)^{\frac{1}{3}}.
\end{aligned}$$

D is the discriminant that determines the nature of the roots of the equation. If $D > 0$, there is one real root and two conjugate complex roots; if $D = 0$, there are real roots of which at least two are equal; if $D < 0$, there are three distinct real roots. In this model, D is negative and there are three distinct real roots. Cardano's formulae for the roots are as follows:

$$\begin{aligned}
root_1 &= -\frac{a_2}{3a_1} + (S + T),\\
root_2 &= -\frac{a_2}{3a_1} - \frac{1}{2}(S + T) + \frac{1}{2}i\sqrt{3}(S - T),\\
root_3 &= -\frac{a_2}{3a_1} - \frac{1}{2}(S + T) - \frac{1}{2}i\sqrt{3}(S - T),
\end{aligned}$$

where $i = \sqrt{-1}$. It turns out for this model that $root_2 < root_3 < root_1$. The optimal solution is $x_1^N = root_3$, and the Nash bargaining solution is $(x_1^N, 1 - x_1^N)$. The solution is only defined for $\beta_1 > 0$ and $\beta_2 > 0$. To make sure the solution is in the bargaining set $\mathcal{B}(\mathcal{U})$, check that $(u_1(x_1^N), u_2(1 - x_1^N)) > (0, 0)$.

Immoral criminals? An experimental study of social preferences among prisoners[*]

Sigbjørn Birkeland d.y. Alexander W. Cappelen
Erik Ø. Sørensen Bertil Tungodden

April 15, 2011

Abstract

Criminal activity has significant costs for society and considerable resources are used on crime prevention. Economists have traditionally focused on how economic opportunities affect criminal behavior and have largely ignored the potential role of social preferences. This paper studies the social preferences of criminals and it is, to our knowledge, the first to do so by conducting an economic lab experiment on a group of prisoners. The main finding in our study is that the prisoners are not immoral in the sense that they are generally unwilling to act on pro-social preferences. Comparing the behavior of the prisoners with the behavior of a benchmark group recruited from a representative sample of the Norwegian population, we find a striking similarity in the importance the two groups attach to pro-social preferences both in strategic and non-strategic situations. We furthermore find little evidence of in-group favoritism in situations where the two groups interact. The pro-social behavior of the prisoners in our experiment clearly contrasts with their anti-social behavior outside the lab. One possible explanation for this cross-situational inconsistency is that behavior in the lab is motivated by different social preferences than behavior outside the lab. The situational inconsistency in behavior could, however, also be seen as suggesting that social preferences are of little importance, compared to circumstances, in explaining criminal behavior.

[*] We would like to thank James Konow and Karl Ove Moene for valuable comments and suggestions. We would also like to thank Harald Aasaune at Bjørgvin Prison and Leif Waage at the The Norwegian Correctional Services for their cooperation, Magne Ludvigsen, Bjørn Ivar Grøttå, Morten Sæthre, and Kristian Øpstad for IT-support, and Henrik Birkeland, Øystein Dvergsdal, Trygve Eiken, Lars Engebretsen, Hans Petter Gramer, Cathrine Gravdal, Ole-Aleksander Greve, Line Halvorsen, Trond Halvorsen, Liv Eline Hetland, Jon Rasmussen, Karoline Salamonsen, Jonas Tungodden, and Fredrik Young-Halvorsen for research assistance. The project was financed by support from the Research Council of Norway, research grant 185831, and the Center for the Study of Mind in Nature, University of Oslo.

1 Introduction

Criminal activity has significant direct costs for society and considerable resources are used on crime prevention. Close to ten million individuals are held in penal institutions around the world (Walmsley, 2009) and in the US alone, 227 billion USD are spent each year to catch, prosecute, and punish offenders (Bureau of Justice Statistics, 2010). It is therefore important to understand what explains criminal behavior.

According to standard economic theory of crime, a person commits a crime if the expected consequences of doing so are better than the expected consequences of any legal alternative. Economists have focused on how differences in legal income opportunities and differences in the expected cost of punishment might explain differences in criminal behavior (Allingham and Sandmo, 1972; Becker, 1968; Dilulio, 1996; Eide, 2000; Levitt, 1997, 2004; Lochner, 2004; McCarthy, 2002). The economic theory of crime has typically ignored how moral considerations might affect the decision to commit a crime, but such considerations could easily be included in the theory as a moral cost of crime (Andvig and Moene, 1990).

Recent experimental studies have documented that many people are motivated by social preferences and often take moral costs into account when they make decisions that have consequences for others (Camerer, 2003). These studies have also shown that there is considerable heterogeneity in pro-social preferences both within and across groups (Henrich, Boyd, Bowles, Camerer, Fehr, and Gintis, 2004). Heterogeneity in pro-social preferences could potentially be important in explaining criminal behavior because crime typically has negative consequences for others (Wikström, 2006; Wikström and Treiber, 2007). More specifically, if people take into account how their actions affect others before they decide whether or not to commit a crime, then the likelihood of a person committing a crime would be decreasing in the importance he attaches to pro-social preferences. Consequently we would expect criminals on average to be less motivated by pro-social preferences than non-criminals.

This paper reports the results from, to our knowledge, the first economic experiment designed to study the social preferences of criminals. The experiment consists of a dictator game and two versions of the trust game: a standard trust game and a trust game with punishment. The dictator game provides us with the classical measure of the importance attached to pro-social preferences in a non-strategic situation. However, as pointed out by Fehr, Naef, and Schmidt (2006), social preferences may differ fundamentally across economic environments, in particular between strategic and non-strategic situations. The trust games allow us to study social preferences in strategic situations (Berg, Dickhaut, and McCabe, 1995).

The prisoners taking part in the experiment were recruited from a medium security prison in Norway. The majority of the prisoners had committed crimes related to drugs, violence, or fraud. In order to have a benchmark with which to compare the behavior of the prisoners, we also included a group of males with the same age distribution as the prisoners, recruited from a representative sample of the Norwegian population.

Group identity has been shown to be important for social preferences in many contexts (Akerlof and Kranton, 2000, 2005; Bernhard, Fehr, and Fischbacher, 2006; Charness, Rigotti, and Rustichini, 2007; Chen and Li, 2009; Tajfel and Rurner, 1979), and in-group favoritism could potentially affect the behavior of the prisoners in the experiment. In addition to single group sessions, where participants only interacted with participants from their own group, we therefore included mixed group sessions where the participants interacted with participants from both groups. This allows us to compare how the prisoners behave when they interact with prisoners and when they interact with participants from the benchmark group. The mixed group sessions also allow us to study whether the benchmark group was prejudiced against the prisoners.

The advantage of a controlled lab experiment in studying the social preferences of prisoners and the benchmark group is that it allows us to compare their behavior in similar choice situations. If the circumstances under which the two groups make their choices are different, which typically would be the case outside the lab, it is not possible to say whether differences in pro-social behavior are a result of differences in circumstances or differences in social preferences. When circumstances are equalized, however, differences in pro-behavior cannot be explained by differences in circumstances.[1]

The main finding in our study is that the prisoners are not immoral in the sense that they are generally unwilling to act on pro-social preferences. On the contrary, we find that the prisoners are highly motivated by pro-social preferences and that there is a striking similarity in the importance the prisoners and the benchmark group attach to pro-social preferences in both non-strategic and strategic situations. This is the case both when the prisoners interact with other prisoners and when they interact with the benchmark group. Thus, we find little evidence of in-group favoritism. Even if our main finding is the similarity in the pro-social preferences of the two groups, we find some interesting differences in how the two groups respond to the punishment option in the trust game. In

[1] The importance of studying social preferences in a setting where differences in circumstances are eliminated can be illustrated by the difference in how the participants in the experiment answered a general trust question and how they answered a specific question about what they believed others would do in the experiment. In a post experimental questionnaire the share of participants who answered that people in general could be trusted was twice as high among the benchmark group as among the prisoners. In contrast, when the participants reported how much they believed the receivers in the trust game would return to the sender, we did not find any difference between the two groups, which suggests that the difference in the answers to the general trust question mainly reflected the fact that the prisoners more often than the benchmark group find themselves in circumstances where they have to be careful in their dealings with people.

particular we find that the prisoners respond less than the benchmark group to a perceived increase in the likelihood of punishment.

We cannot exclude the possibility that the prisoners' behavior in the experiment was affected by the fact that they were imprisoned. The prisoners could, for example, have been more affected by the scrutiny of the experimental situation than the benchmark group (Levitt and List, 2007). Special care was, however, taken to reduce this effect and to make the lab experience as similar as possible for the two groups. In particular we made sure that no prison guards were present in the lab during the experiment and at the beginning of the experiment we explained to the participants the procedures ensuring that it was impossible for the experimenters, or anyone else, to link subjects to individual choices. The experiment was furthermore highly incentivized, with an average payment, excluding show-up fee, of 482 NOK (approximately 85 USD).

Our main result could be seen as providing support to the claim made in Becker (1968), that criminals do not differ from non-criminals with respect to their basic motivation and that differences in the costs and benefits from crime are the main explanation for differences in criminal activity. Given this interpretation, the striking cross-situational differences in the pro-social behavior of the prisoners inside and outside the lab could reflect that the prisoners face the same circumstances as the benchmark group in the lab, while the circumstances they face outside the lab are very different. An alternative interpretation of our main results, however, is that there is weak cross-situational consistency in social preferences, and that behavior in the lab is motivated by different social preferences than behavior outside the lab (Levitt and List, 2007).

Section 2 and Section 3 present the sampling procedure and the experimental design. Section 4 and Section 5 analyze pro-social preferences in the dictator game and the trust game, respectively. Section 6 discusses some implications of our findings.

2 Sample

We conducted 12 sessions, which on average lasted 90 minutes, with a total of 360 participants during the period from June 2007 to April 2009. Four sessions, with a total of 207 participants, were mixed sessions in which the prisoners and the benchmark group interacted, and eight sessions, with a total of 153 participants, were single group sessions in which the participants only interacted with participants from their own group. No individual participated in more than one session.

The 187 prisoners who participated in the experiment were all male inmates of Bjørgvin Prison, a medium security prison located outside the city of Bergen, Norway. The prisoners were invited to participate in the experiment at a meeting some days in advance of each session. At the meeting we also handed out written invitations in which we explained that the experiment was voluntary, that participants would not be asked to reveal any personal information, and that any information gathered in the experiment would be anonymous. They were furthermore informed that in addition to a show-up fee they could earn extra money during the experiment, that all earnings would be paid in cash immediately after the experiment, and that they did not have to report their earnings from the experiment to the prison authorities. At Bjørgvin Prison, the experiment was conducted in a mobile computer lab that was set up in the prison gymnasium and no prison guards were present in the lab during the experiment.

The other group of participants consisted of 173 males selected randomly from the population living in the 27 basic statistical units closest to the Norwegian School of Economics and Business Administration (NHH) in Bergen.[2] These basic statistical units include parts of the second largest city in Norway as well as a less populated rural area, and the population is close to the national average with respect to the distribution of income, education and occupation. The inmates at Bjørgvin prison are on average younger than the general population, and we stratified the invitations so that the age profile of the benchmark group was approximately the same as for the prisoners.[3] Table 1 reports the characteristics of the two groups based on self-reported age, education and work experience. We observe that the two groups are very similar with respect to age and work experience, but that a somewhat higher share of the benchmark group has completed secondary education.

[Table 1 about here.]

The benchmark group received an invitation letter similar to the one received by the prisoners and they were given the same instructions during the experiment. These participants conducted the experiment at NHH, where we set up a computer lab of the same type as the one used in the gymnasium at Bjørgvin Prison.[4]

3 Design

The experiment consisted of two parts: a dictator game and a version of the trust game. It was conducted using a web-based interface and was double blind so

[2] A basic statistical unit is the smallest geographical unit used by Statistics Norway.

[3] The selection procedure was approved by the Norwegian Social Science Data Services ("Norsk samfunnsvitenskaplig datatjeneste") and the Norwegian Public Register ("Norsk Folkeregister").

[4] To compensate the benchmark group for the additional time and costs incurred by this group in order to come to the lab, the show-up fee for the benchmark group, 300 NOK, was higher than the show-up fee for the prisoners, 100 NOK. The participants were not informed about the other group's show-up fee.

that neither subjects nor experimenters could associate decisions with particular subjects. No information about the outcome of the dictator game was given to the participants before both parts of the experiment were completed.

At the beginning of the experiment, all participants were informed about the rules of conduct and given a description of how the experiment would proceed.[5] Instructions were given by the experimenter and on the computer screens. To prevent participants with poor reading skills from misunderstanding the written instructions, it was possible to listen to a pre-recorded version of the instructions using headsets available to all participants. In all the mixed sessions, the participants were told the location of the other participant, in Bjørgvin Prison or at NHH, and given a short description of how the participants in the other location had been recruited, that the participants at Bjørgvin Prison were male inmates at a medium security prison, and that the participants at NHH were males recruited from the general population.

In the dictator game, the participants were asked to divide an endowment of money between themselves and another participant. Each participant made this decision in two situations and was a recipient in two other situations. In each situation they were matched with a different participant and they were not informed about the outcome in the two situations where they were a recipient before at the end of the session.

The endowment to be distributed by the dictator in the mixed session was 1000 NOK (approximately 175 USD). In the single group session each participant was a dictator in one situation with an endowment of 1000 NOK and in one situation with an endowment of 500 NOK. The dictators could give the other participant six alternative shares of the endowment: 0 percent, 20 percent, 40 percent, 60 percent, 80 percent or 100 percent.

Immediately after the dictator game, the participants took part in one of two versions of a trust game: a standard trust game or a trust game with a punishment option. Each participant was involved in four trust situations; first in two situations as a sender and then in two situations as a receiver, and in each situation they were given an endowment of 400 NOK (approximately 70 USD). They did not receive any information about the outcome of the first two situations before they decided how much to return in the situations in which they were receivers. In each situation they were matched with a different participant. In the mixed sessions they were matched with one participant from each location, both as a sender and as a receiver, and the participants knew the location of the other participant when they made their decisions.

In the standard trust game the senders were given the opportunity to send up to 200 NOK, choosing among six alternative shares of this amount: 0 percent, 20 percent, 40 percent, 60 percent, 80 percent and 100 percent. The amount sent was multiplied by a factor of three so that the receiver received three times

[5] Complete instructions can be downloaded from http://sites.google.com/site/sameos/.

the sent amount. Before the sender made his choice, he was informed that the receiver could return six alternative shares of the received amount: 0 percent, 20 percent, 40 percent, 60 percent, 80 percent or 100 percent, and was asked to report what he believed the probability was that the receiver would return each of the alternative shares.[6] The sender thus had to reflect on how the receiver would respond to his decisions before he decided what to do.

When the sender had decided how much to send, the receiver was informed of how much he had received. The receiver then had to decide what share of the received amount (0 percent, 20 percent, 40 percent, 60 percent, 80 percent or 100 percent) he wanted to return to the other participant.

The trust game with a punishment option was identical to the standard trust game except that the sender had the option to punish the other participant. In the mixed sessions, the sender could choose to reduce the other participants's payoff by 100 NOK or 200 NOK at a low cost to himself (0.25 NOK per 1 NOK reduction). In the single group sessions the punishment cost for the sender was low in one of the situations and high in the other (1 NOK per 1 NOK reduction).

Before the receiver made his choice of how much to return he was asked, for each possible return amount, to report what he believed the probability was that the sender would choose to reduce his payment by 0 NOK, 100 NOK or 200 NOK.[7] After the receiver had decided how much to return, the sender decided whether he wanted to punish the receiver by reducing his payment by 100 NOK or 200 NOK at a cost to himself.

Throughout the experiment, after having made a decision the participants were immediately shown the consequences of their decision and then asked to either confirm or revise it. At the end of each part of the experiment, they were again given the opportunity to revise all of their decisions in that part, and then asked to make a final confirmation of their decisions.

At the end of the experiment, one of the eight situations each participant was involved in, four dictator game situations and four trust game situations, was randomly drawn and the participant received his earnings from this situation in addition to the show-up fee.[8] The average earnings, excluding the show-up fee, was 482 NOK (approximately 85 USD). Special care was taken so that the payment procedure ensured anonymity. The computer assigned a payment code to each of the participants, and a group of assistants, who were not present in the lab during the experiment, prepared envelopes containing the payments

[6] In the mixed sessions, the sender was asked to report these beliefs both when the receiver was a prisoner and when the receiver was from the benchmark group.

[7] In the mixed sessions the participants answered these questions both when the sender was a prisoner and when the sender was from the benchmark group, and in the single group sessions they answered these questions for high and low punishment cost.

[8] Due to a computer error, five participants in the benchmark group only made one choice as a dictator and one choice as a sender in the trust game. Total number of observations from this group is therefore 341 and not 346.

corresponding to each payment code. The assistants also made sure that it was impossible to identify the amount of money by simply looking at the envelope. After bringing the envelopes to the lab, the assistants immediately left and the envelopes were handed out in accordance with the payment codes. The sequence of events in the two versions of the experiment is summarized in Table 2.

[Table 2 about here.]

4 Social preferences in non-strategic situations

The distributive situation in the dictator game has two important characteristics that limit the possible motives the dictator may have for sharing. First, the situation is non-strategic in the sense that the other participant is unable to respond to the decision made by the dictator, which implies that sharing cannot be motivated by self-interest. Second, the dictator does not respond to decisions made by the other participant, which implies that sharing cannot be motivated by reciprocity. Sharing in the dictator game could, however, be motivated by both inequality aversion and by altruism.

The upper left panel in Figure 1 provides a histogram of the share given in situations where prisoners are matched with other prisoners, where we observe that the large majority gives something to the other participant. Some prisoners give nothing to the other participant, but the modal choice is to give 40 percent of the endowment.[9] From Table 3 we observe that the prisoners give on average 36.2 percent to the other participant, which is more than commonly reported for experiments conducted with students (Camerer, 2003).

[Figure 1 about here.]

[Table 3 about here.]

The dictators were informed about the location of the other participant and this information could potentially affect their sharing behavior. Prisoners could, for example, be more willing to act on pro-social preferences when they were matched with other prisoners than when they were matched with a participant from the benchmark group. Comparing the upper left and the upper right panels in Figure 1 we observe, however, that the distribution of shares given is very similar in the two types of situations. From Table 3 we observe that the prisoners on average give slightly more to other prisoners than they give to participants from the benchmark group, but the difference is not significant ($p = 0.134$).

[9] There is no significant difference in the average share given when the endowment is 500 NOK and 1000 NOK. The prisoners gave 2.7 percentage points less with the high endowment than with the low endowment ($p = 0.242$) and the benchmark group gave 1.2 percentage points more ($p = 0.420$).

Comparing the upper panels and the lower panels in Figure 1 we observe a striking similarity in the distribution of shares given for the prisoners and the benchmark group. This impression is confirmed by Table 3 where we find no significant difference in the average share given ($p = 0.273$). This similarity in the share given also holds when we look separately at how much each of the two groups gives in situations where they are matched with prisoners ($p = 0.426$), and in situations where they are matched with the benchmark group ($p = 0.601$).

Table 4 reports a regression on share given where we control for age, education and work experience. The coefficient for the dictator being a prisoner is insignificant, which confirms the impression that there is no difference in the weight the two groups attach to pro-social preferences. We also observe that both groups give somewhat more when the recipient is a prisoner than they do when the recipient is from the benchmark group, but this is only significant for the benchmark group.

In sum, we find that the prisoners are highly motivated by pro-social preferences in the dictator game and that there are no differences in the sharing behavior of the prisoners and the benchmark group. We therefore conclude that prisoners are not characterized by an unwillingness to act on pro-social preferences in non-strategic situations, neither in meetings with other prisoners or with participants from the general population.

5 Social preferences in strategic situations

We now turn to the trust game, which allows us to study the participants' social preferences in strategic situations. Since the participants in the trust game respond to decisions made by other participants and have to take into account how other participants respond to their decisions, this game introduces motives that are not present in the dictator game.

The decision to send does not provide a clean measure of pro-social behavior since it also is affected by beliefs about the other participant's behavior, and our analysis therefore focuses on the return decision and on the decision to punish in the trust game with a punishment option. We observe, however, from Table 5, that the average share sent by the prisoners and by the benchmark group is strikingly similar if we look at all the situations, 62.2 percent versus 64.1 percent. Looking only at the standard trust game, we observe that prisoners send somewhat less than the benchmark group, in particular when the receiver is a prisoner, but this difference is not statistically significant ($p = 0.183$). The share sent in the standard trust game is often interpreted as a measure of trust (Fehr, 2009), and this result therefore suggests that the two groups are equally trusting when they make choices in the same circumstances.

[Table 5 about here.]

5.1 Share returned

Table 6 reports the average share returned in the trust game for the prisoners and the benchmark group. If we look at the average across all return decisions, the prisoners return 38.6 percent of the received amount and the benchmark group returns 41.3 percent. However, since the decision to return in the trust game with a punishment option might be motivated by a desire to avoid punishment, the average share returned does not provide a clean measure of pro-social motivation.

In the standard trust game only pro-social preferences can motivate the receiver to return a share of the received money and we observe that the prisoners in these situations return close to one third of the money they receive. From Table 6 we observe that there is no important difference in the average share returned when the sender is a prisoner and when the sender is from the benchmark group ($p = 0.179$). Prisoners are thus motivated by pro-social preferences also in situations where they respond to others decisions.

Comparing the average share returned by the prisoners and the benchmark group in the standard trust game, we observe from Table 6 that the benchmark group returns a higher share, but the difference is not statistically significant, ($p = 0.157$). Since self-interest cannot explain a positive amount returned, this result suggests that both groups are equally motivated by pro-social preferences in their interaction with participants from the benchmark group. When the sender is a prisoner, however, we observe that the benchmark group return somewhat more than the prisoners, ($p < 0.001$).

The trust game with a punishment option introduces an additional motive of avoiding punishment, a motive that in itself should make the participants more motivated to return a high share. We observe from Table 6 that the prisoners return a higher share when there is a punishment option than they do in the standard trust game, in particular when the cost of punishment is low. For the benchmark group, we observe no systematic effect of the punishment option, which may reflect that the presence of a punishment option crowds out the moral motivation for the benchmark group.

[Table 6 about here.]

Is the return decision in the trust game motivated by the same pro-social preferences that motivated sharing in the standard dictator game? To address this question, and to study the role of reciprocity and punishment, Table 7a reports a regression of the share returned by the prisoners in all the return decisions.

To capture the pro-social preferences that motivate the participants in the dictator game we calculate the amount that each participant has to return in order to achieve the distribution he selected as a dictator. We define the variable 'Dictator' as the maximum of this number and zero. [10] This variable is the

[10] A similar approach is used in Ashraf, Bohnet, and Piankov (2006); Cappelen, Nygaard, Sørensen, and Tungodden (2010).

amount the participant would return if he wanted the distribution in the trust game to be as close as possible to the distribution he chose when he was a dictator, taking into account that it is impossible to return a negative amount.

From column (5) in Table 7a we observe that the Dictator variable has a large and significant effect on how much the prisoners return, which suggests that the pro-social preferences salient in the non-strategic dictator game also are important motives in the strategic trust game. Table 7b reports the regression of share returned for the benchmark group and we find that the cross-situational consistency in pro-social preferences also holds for the benchmark group.

A large body of evidence has shown that many people are willing to reward kind actions even at a cost to themselves (Fehr and Gachter, 2000; Falk and Fischbacker, 2006). Reciprocity could potentially be important for the receivers in the trust game since they are placed in a distributive situation where the sender may have acted kindly by sending an amount. If the receivers are motivated by a desire to reciprocate, we would expect the share returned to be increasing in the share sent. From column (5) in Table 7a we observe, however, that the share returned by the prisoners is not increasing in the share sent, which suggests that reciprocity is not an important motive for them in this situation. Comparing with Table 7b we observe that the same holds for the benchmark group.

A desire to avoid punishment could affect the return decision in the situations where the sender had a punishment option. In the regression we look at the effect of the two different punishment options, Low cost and High cost, and the effect of a marginal increase in the belief that the sender will use the punishment option. From column (5) in Table 7a we observe that the existence of a punishment option has a positive effect on the share returned by the prisoners, but this effect is only significant when the cost is low. We follow the approach of Falk, Meier, and Zehnder (2011) and use the average expected punishment as a proxy for the belief that the sender is likely to use the punishment option and we observe that the coefficient for this 'Belief' variable is small and insignificant for the prisoners. This suggests that expectations about punishment are of little importance when the prisoners decide how much to return.

In contrast to what we find for the prisoners, the benchmark group responds to the punishment option by reducing how much they return. This suggests that pro-social motivation of the benchmark group is crowded out by the threat of punishment. The benchmark group also differ from the prisoners in how they respond to an increased likelihood that the other participant will use the punishment option. Column (5) in Table 7b shows that Belief has a large and significant effect on the share returned for the benchmark group ($p = 0.001$).

From Table 7a and Table 7b we observe that the indicator for sender being a prisoner is not significant for either group and the interaction terms between sender being a prisoner and Dictator, Share sent and Beliefs are also insignificant. In line with the results from the dictator game, we thus find no evidence of in-group favoritism.

[Table 7 about here.]

In sum, the high average share returned by the receivers in the standard trust game shows that both the prisoners and the benchmark group are highly motivated by pro-social preferences. The return decision seems to a large extent to be motivated by the same pro-social preferences that motivated sharing in the dictator game. This suggests that social preferences are consistent across very different economic environments within an experimental setting. Reciprocity does not seem to be an important motive for either of the two groups. The two groups differ, however, in how they respond to the punishment option. The existence of of a punishment option increases the share returned among the prisoners, but it has a large negative effect on the share returned among the benchmark group. For both groups, the share returned is higher the more likely the participant think it is that the punishment option will be used, but this effect is only statistically significant for the benchmark group. In line with what we found in the dictator game, there is no evidence of in-group favoritism.

5.2 Punishment

In the previous section we found that reciprocity was of little importance for both groups when they decided how much of the received money they would return to the sender. The trust game with a punishment option allows us to study whether reciprocity is more important in the decision to punish than in the return decision. The punishment option is also interesting because it creates a situation where inequality aversion might conflict with altruism. Punishment can, when the cost of punishment is low, equalize the final income distribution and inequality averse participants therefore have a motive to punish.[11] Altruism would, however, be a reason not to punish because punishment reduces the income of the other participant.

None of the participants choose to punish the receiver when the cost of punishment is high. From Table 8, we observe, however, that when the cost of punishment is low, 21.5 percent of the prisoners choose to punish and their average punishment is 34 NOK. For the benchmark group the corresponding numbers are 23.8 percent and 39.2 NOK. Even if there is no significant difference in the average levels of punishment, the two groups differ with regard to who they punish.

[Table 8 about here.]

Table 9 reports marginal effects for a probit regression on the decision to punish when the punishment cost is low and it allows us to study how different

[11] This is not the case when the cost of punishment is high, in which case the level of inequality is unaffected by the level of punishment.

motives affect the decision to punish.[12] If reciprocity is an important motive in the punishment decision we would expect the participants to be more likely to punish when the share returned is low. From Table 9 we observe that this indeed is the case. Both prisoners and the benchmark group are significantly less likely to punish if the receiver returns a high share of the received amount ($p < 0.001$). Reciprocity is thus an important motive for both groups when they decide whether or not to punish.

In the dictator game, inequality aversion and altruism are both motives for sharing. With respect to punishment, however, these motives pull in opposite directions. The correlation between pro-social behavior in the dictator game and punishment in the trust game will therefore depend on the relative importance of these two motives in the dictator decision. In Table 9 "Dictator" is defined as the punishment that is required in order to come as close as possible to the distribution the participant selected as a dictator. We observe that the Dictator variable has a large negative effect on punishment for the prisoners ($p = 0.031$). In contrast, the Dictator variable has no effect on punishment for the benchmark group ($p = 0.838$). There are two plausible interpretations of this difference between the two groups. First, it could be seen as suggesting that the prisoners place less weight on inequality aversion relative to altruism than the benchmark group. However, it could also be seen as suggesting that the prisoners view punishment as a more anti-social act than the benchmark group, and that they therefore are more reluctant to punish if they are highly motivated by pro-social preferences.

The prisoners are less likely to punish other prisoners than participants from the benchmark group, while the benchmark group is more likely to punish prisoners than members of their own group, but when we control for background variables and other motives these effects are not significant ($p = 0.103$ and $p = 0.332$ respectively).

[Table 9 about here.]

In sum, we find that the prisoners are not characterized by being more willing to punish others than the benchmark group. In contrast to what we find for the decision to return, reciprocity is an important motive in the punishment decision for both groups. For the prisoners there is strong negative correlation between pro-social behavior in the dictator game and the willingness to punish, but there is no such correlation for the benchmark group. One interpretation of this result is that prisoners perceive punishment as a more anti-social act than the benchmark group.

[12] OLS regressions give very similar results.

6 Conclusion

The results from the experiment presented in this paper suggest that prisoners are not immoral in the sense that they are characterized by a general unwillingness to act on pro-social preferences. On the contrary, we find no major differences in the pro-social preferences of the prisoners and a benchmark group recruited from the general population when they face the same circumstances. Importantly, this result is not driven by in-group favoritism among the prisoners.

The main differences we find between the prisoners and the benchmark group in our experiment are related to the punishment option in the trust game. We find that prisoners increase their pro-social behavior when the punishment option is introduced, but that they do not respond to an increase in expected punishment. The benchmark group, in contrast, decreases their pro-social behavior when the punishment option is introduced, but responds strongly to an increase in expected punishment. For the prisoners we also find that there is a strong negative correlation between pro-social behavior in the dictator game and their willingness to punish, but we find no such correlation for the benchmark group. This result might suggest that prisoners view punishment as a more anti-social act than the benchmark group.

If social preferences were important in explaining criminal behavior we would expect prisoners on average to be less motivated by pro-social preferences than the benchmark group in our experiment. In contrast, we find a striking similarity between pro-social motivation of the two groups in the experiment. We consider two plausible interpretations of this result. One interpretation is that there is weak cross-situational consistency in social preferences and that our results therefore cannot be extrapolated from the experimental setting to the world outside the lab. Interestingly, however, we find considerable consistency in pro-social behavior across different situations in our experiment, but we cannot exclude the possibility that such consistency does not apply when we move out of the lab.

An alternative interpretation of our results is that social preferences are of little importance in explaining criminal behavior and that differences in criminal behavior primarily are a result of differences in circumstances. This interpretation evokes a puzzle in light of the fact that pro-social preferences seem to be important in many other contexts where people make decisions. One explanation for this puzzle could be that criminal behavior to a large extent is caused by lack of self-control, a view common among criminologists (Gottfredson and Hirschi, 1990). If this is the case, criminals can be highly motivated by pro-social preferences in situations where they have self-control, but sometimes be unable to act on these preferences because they lose their self-control. An interesting avenue for further research is to investigate the interaction between social preferences and self-control in explaining criminal behavior.

A related issue for further research is whether there are systematic differences in the social preferences of different types of criminals. In order to secure the complete anonymity of the prisoners we did not ask them about what type of crime they had committed. We can therefore not rule out that particular groups of criminals are characterized by attaching little importance to pro-social preferences. It could, for example, be the case that prisoners who are convicted of crimes that require pre-meditation, such as certain types of white-collar crime, are characterized by being less motivated by pro-social preferences than other criminals.

References

Akerlof, George A. and Rachel E. Kranton (2000). "Economics and identity," *Quarterly Journal of Economics*, 115(3): 715–753.

Akerlof, George A. and Rachel E. Kranton (2005). "Identity and the economics of organizations," *Journal of Economic Perspectives*, 19(1): 9–32.

Allingham, Michael G. and Agnar Sandmo (1972). "Income tax evasion: A theoretical analysis," *Journal of Public Economics*, 1(3-4): 323–338.

Andvig, Jens Chr. and Karl Ove Moene (1990). "How corruption may corrupt," *Journal of Economic Behavior and Organization*, 13(1): 63–76.

Ashraf, Nava, Iris Bohnet, and Nikita Piankov (2006). "Decomposing trust and trustworthiness," *Experimental Economics*, 9(3): 193–208.

Becker, Gary S. (1968). "Crime and punishment: An economic approach," *The Journal of Political Economy*, 76(2): 169–217.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995). "Trust, reciprocity, and social history," *Games and Economic Behavior*, 10(1): 122–142.

Bernhard, Helen, Ernst Fehr, and Urs Fischbacher (2006). "Group affiliation and altruistic norm enforcement," *American Economic Review*, 96(2): 217–21.

Bureau of Justice Statistics (2010). "Justice expenditure and employment extracts, 2007," http://bjs.ojp.usdoj.gov/index.cfm?ty=pbdetail&iid=2315.

Camerer, Colin F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton, NJ: Princeton University Press.

Cappelen, Alexander, Knut Nygaard, Erik Ø. Sørensen, and Bertil Tungodden (2010). "Efficiency, equality and reciprocity in social preferences: A comparison of students and a representative population." Discussion paper SAM 28/10, Department of Economics, Norwegian School of Economics and Business Administration.

Charness, Gary, Luca Rigotti, and Aldo Rustichini (2007). “Individual behavior and group membership,” *American Economic Review*, 97(4): 1340–52.

Chen, Yan and Sherry Xin Li (2009). “Group identity and social preferences,” *American Economic Review*, 99(1): 431–457.

Dilulio, John J. (1996). “Help wanted: economists, crime and public policy,” *Journal of Economic Perspectives*, 10(1): 3–24.

Eide, Erling (2000). “Economics of criminal behavior,” in Boudewijn Bouckaert and Gerrit De Geest (eds.), “Encyclopedia of Law and Economics,” volume 5, Edward Elgar, pp. 345–389.

Falk, Armin and Urs Fischbacker (2006). “A theory of reciprocity,” *Games and Economic Behavior*, 54(2): 293–315.

Falk, Armin, Stephan Meier, and Christian Zehnder (2011). “Did we overestimate the role of social preferences? the case of self-selected student samples,” *IZA Discussion Papers*, 5475.

Fehr, Ernst (2009). “On the economics and biology of trust,” *Journal of the European Economic Association*, 7(2-3): 235–266.

Fehr, Ernst and S. Gachter (2000). “Fairness and retaliation: The economics of reciprocity,” *Journal of Economic Perspectives*, 14(3): 159–181.

Fehr, Ernst, Michael Naef, and Klaus M. Schmidt (2006). “Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment,” *American Economic Review*, 96(5): 1912–1917.

Gottfredson, Michael R. and Travis Hirschi (1990). *A General Theory of Crime*, Stanford University Press.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis (eds.) (2004). *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford, UK: Oxford University Press.

Levitt, Steven D. (1997). “Using electoral cycles in police hiring to estimate the effects of police on crime:,” *American Economics Review*, 87(3): 270–290.

Levitt, Steven D. (2004). “Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not,” *Journal of Economic Perspectives*, 18(1): 163–190.

Levitt, Steven D. and John A. List (2007). “What do laboratory experiments measuring social preferences reveal about the real world?” *The Journal of Economic Perspectives,*, 21(2): 153–174.

Lochner, Lance (2004). “Education, work, and crime: A human capital approach,” *International Economic Review*, 45(3): 811–843.

McCarthy, Bill (2002). “New economics of sociological criminology,” *Annual Review of Sociology*, 28: 417–442.

Tajfel, Henri and John Rurner (1979). “An integrative theory of intergroup conflict,” in Stephen Worchel and William Austin (eds.), “The Psychology of Intergroup Relations,” Chicago: Nelson-Hall, pp. 7–24.

Walmsley, Roy (2009). “World prison population list, eighth edition.” International Centre for Prison Studies.

Wikström, Per-Olof (2006). “Linking individual, setting and acts of crime. situational mechanisms and the explanation of crime,” in Per-Olof Wikström and Robert Sampson (eds.), “The Explanation of Crime: Contexts, Mechanisms and Development,” Cambridge University Press, pp. 61–107.

Wikström, Per-Olof and Kyle Treiber (2007). “The role of self-control in crime causation: Beyond Gottfredson and Hirschi’s general theory of crime,” *European Journal of Criminology*, 4(2): 237–264.

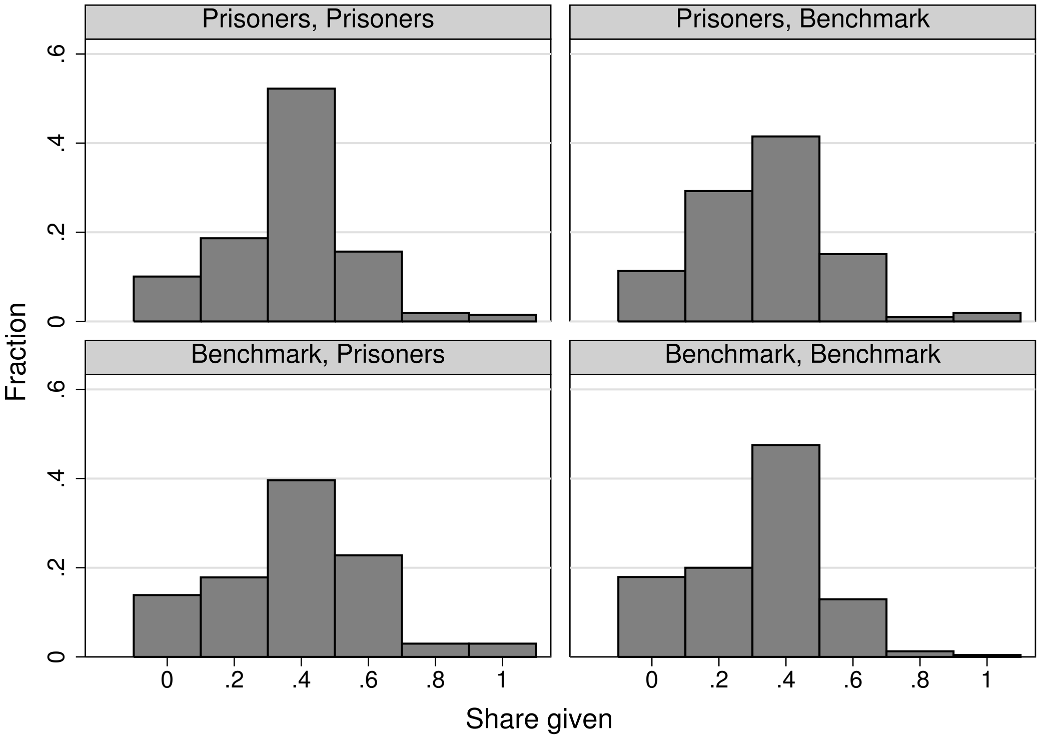


Figure 1: Histograms of share given

Note: Share given is the share of the endowment given to the other participant in the dictator game. "A, B" should be read as the situations where a participant from subject group A decides how much to give to a participant from subject group B.

Table 1: Sample characteristics

	Prisoners	Benchmark
Above 25 years old	0.764	0.774
	(0.031)	(0.032)
Completed secondary school	0.631	0.879
	(0.035)	(0.025)
At least five years work experience	0.727	0.722
	(0.037)	(0.034)
n	187	173

Note: Self-reported age, education and work experience. Standard errors in parentheses.

Table 2: Sequence of events

Stage of experiment	T1	T2
1. Decisions to share in dictator game	yes	yes
2. Beliefs about share returned in trust game	yes	yes
3. Decisions to send in trust game	yes	yes
4. Beliefs about punishment in trust game	no	yes
5. Decisions to return in trust game	yes	yes
6. Decisions to punish in trust game	no	yes
7. One situation drawn for payment	yes	yes

Note: Sequence of events in the experiment. T1: the standard trust game; T2: trust game with a punishment option.

Table 3: Average share given in the dictator game

	Receiver		
Sender	Prisoner	Benchmark	All
Prisoner	0.370	0.342	0.362
	(0.014)	(0.020)	(0.013)
n	268	106	374
Benchmark	0.384	0.322	0.340
	(0.023)	(0.015)	(0.015)
n	101	240	341

Note: The table reports average share given in the dictator game, standard errors (in parentheses) are corrected for clustering on individuals, and n is the number of observations.

Table 4: Regressions of share given

	Prisoner	Benchmark	All
Dictator is prisoner			0.008
			(0.022)
Other participant is prisoner	0.032	0.060	0.046
	(0.019)	(0.021)	(0.014)
Age	0.010	0.102	0.037
	(0.032)	(0.057)	(0.028)
Education	0.020	0.011	0.020
	(0.027)	(0.040)	(0.023)
Work experience	0.040	-0.030	0.024
	(0.029)	(0.053)	(0.026)
Constant	0.290	0.254	0.263
	(0.035)	(0.047)	(0.032)
Observations	374	341	715
R^2	0.019	0.046	0.028

Note: Regression of share given in the dictator game on background variables. Age, Education and Work experience are indicator variables taking the value one when age is above 25 years, when secondary education is completed, and when work experience is at least five years. The left column is based on all situations where the dictator is a prisoner, the middle column is based on all situations where a participant from the benchmark group is a dictator, and the right column is based on all situations. Standard errors (in parentheses) are corrected for clustering on individuals.

Table 5: Share sent

		T1		T2-low		T2-high	
		Recipient		Recipient		Recipient	
Sender	All	Prisoner	Benchmark	Prisoner	Benchmark	Prisoner	Benchmark
Prisoner	0.622	0.586	0.600	0.663	0.682	0.586	
	(0.024)	(0.039)	(0.047)	(0.034)	(0.043)	(0.056)	
n	374	128	50	98	56	42	
Benchmark	0.641	0.735	0.664	0.600	0.636		0.541
	(0.025)	(0.050)	(0.043)	(0.042)	(0.033)		(0.058)
n	340	43	107	58	95		37

Note: Average share sent in the standard trust game (T1) and the trust game with punishment (T2). “Low” and “High” refer to the price of punishment and n is the number of observations. Standard errors (in parentheses) are corrected for clustering on individuals.

Table 6: Share returned

		T1 Sender		T2-low Sender		T2-high Sender	
Receiver	All	Prisoner	Benchmark	Prisoner	Benchmark	Prisoner	Benchmark
Prisoner	0.386	0.335	0.309	0.443	0.458	0.395	
	(0.018)	(0.028)	(0.040)	(0.025)	(0.038)	(0.042)	
n	343	113	46	94	52	38	
Benchmark	0.413	0.512	0.382	0.449	0.420		0.300
	(0.021)	(0.042)	(0.032)	(0.039)	(0.030)		(0.049)
n	320	41	99	57	89		34

Note: Average of share returned in the standard trust game (T1) and in the trust game with punishment (T2). "Low" and "High" refer to the cost of punishment and n is the number of situations with a positive sent amount. Standard errors (in parentheses) are corrected for clustering on individuals.

Table 7: Regressions of share returned

(a) Prisoners

	(1)	(2)	(3)	(4)	(5)
Dictator	0.574				0.674
	(0.113)				(0.127)
Share sent		0.056			-0.149
		(0.057)			(0.084)
Beliefs			0.013		0.051
			(0.064)		(0.065)
Sender is prisoner				0.007	-0.040
				(0.026)	(0.079)
Dictator X sender is prisoner	-0.013				-0.084
	(0.104)				(0.140)
Share sent X sender is prisoner		0.023			0.102
		(0.036)			(0.104)
Beliefs X sender is prisoner			-0.017		-0.041
			(0.043)		(0.060)
Low cost	0.117	0.123	0.120	0.122	0.092
	(0.033)	(0.037)	(0.055)	(0.037)	(0.046)
High cost	0.079	0.074	0.078	0.073	0.062
	(0.044)	(0.048)	(0.061)	(0.048)	(0.053)
Age	-0.014	-0.013	-0.017	-0.017	-0.019
	(0.039)	(0.044)	(0.044)	(0.044)	(0.039)
Education	0.036	0.041	0.038	0.039	0.037
	(0.031)	(0.036)	(0.037)	(0.036)	(0.031)
Work experience	0.012	0.033	0.040	0.039	0.014
	(0.033)	(0.039)	(0.039)	(0.039)	(0.033)
Constant	0.202	0.237	0.287	0.281	0.278
	(0.048)	(0.058)	(0.049)	(0.054)	(0.083)
Observations	343	343	343	343	343
R^2	0.206	0.066	0.058	0.058	0.215

Continued on next page.

Table 7: Regressions of share returned (continued)

(b) Benchmark

	(6)	(7)	(8)	(9)	(10)
Dictator	0.645				0.783
	(0.090)				(0.122)
Share sent		-0.009			-0.137
		(0.054)			(0.070)
Beliefs			0.188		0.204
			(0.081)		(0.062)
Sender is prisoner				0.071	0.124
				(0.027)	(0.081)
Dictator X sender is prisoner	0.081				-0.109
	(0.120)				(0.220)
Share sent X sender is prisoner		0.056			-0.047
		(0.035)			(0.122)
Beliefs X sender is prisoner			0.004		-0.046
			(0.041)		(0.060)
Low cost	0.044	0.027	-0.118	0.025	-0.109
	(0.035)	(0.042)	(0.084)	(0.042)	(0.060)
High cost	-0.042	-0.085	-0.234	-0.074	-0.183
	(0.053)	(0.057)	(0.077)	(0.056)	(0.060)
Age	-0.082	-0.034	-0.060	-0.033	-0.119
	(0.057)	(0.066)	(0.065)	(0.065)	(0.052)
Education	-0.051	0.000	0.021	-0.003	-0.027
	(0.049)	(0.051)	(0.048)	(0.049)	(0.045)
Work experience	0.210	0.166	0.152	0.166	0.202
	(0.051)	(0.058)	(0.056)	(0.056)	(0.046)
Constant	0.234	0.310	0.328	0.296	0.302
	(0.052)	(0.062)	(0.059)	(0.059)	(0.061)
Observations	320	320	320	320	320
R^2	0.275	0.076	0.100	0.084	0.338

Note: Regression of share returned in the trust game. “Dictator” is the share the participant has to return in order to come as close as possible to the distribution chosen in the dictator game. “Beliefs” is the average expected punishment measured in units of 100 NOK. “Low cost” and “high cost” refer to the price of punishment in the trust game with a punishment option. “Sender is prisoner” is an indicator variable taking the value one if the sender is a prisoner, “Age” is an indicator variable taking the value one if age is above 25 years, “Education” is an indicator variable taking the value one if secondary school is completed, and “Work experience” is an indicator variable taking the value one if work experience is at least five years. Standard errors (in parentheses) are corrected for clustering on individuals.

Table 8: Average punishment

	Share that punishes			Punishment in NOK		
	Receiver			Receiver		
Sender	All	Prisoner	Benchmark	All	Prisoner	Benchmark
Prisoners	0.215	0.160	0.309	34.23	23.40	52.73
	(0.038)	(0.038)	(0.063)	(6.43)	(5.94)	(11.22)
n	149	94	55	149	94	55
Benchmark	0.238	0.278	0.213	39.16	42.59	37.08
	(0.040)	(0.061)	(0.044)	(6.90)	(10.06)	(7.89)
n	143	54	89	143	54	89

Note: Share of participants who punish and average punishment in NOK by receiver type. n is the number of situations with low price of punishment and where a positive amount was sent. Standard errors (in parentheses) corrected for clustering on individuals.

Table 9: Regression of punishment

A: Prisoners	(1)	(2)	(3)	(4)
Dictator	-0.494			-0.685
	(0.237)			(0.241)
Share returned		-0.374		-0.586
		(0.112)		(0.159)
Receiver is prisoner			-0.136	-0.203
			(0.064)	(0.148)
Share returned X Receiver is prisoner		-0.221		0.146
		(0.090)		(0.216)
Age	0.123	0.167	0.131	0.132
	(0.068)	(0.063)	(0.069)	(0.059)
Education	0.140	0.147	0.147	0.119
	(0.069)	(0.071)	(0.069)	(0.070)
Work experience	-0.181	-0.195	-0.203	-0.159
	(0.105)	(0.108)	(0.103)	(0.101)
Observations	148	148	148	148
log likelihood	-69.803	-63.336	-70.597	-57.653

B: Benchmark	(1)	(2)	(3)	(4)
Dictator	0.258			-0.046
	(0.256)			(0.226)
Share returned		-0.872		-0.804
		(0.173)		(0.173)
Receiver is prisoner			0.059	0.105
			(0.068)	(0.112)
Share returned X Receiver is prisoner		0.155		-0.067
		(0.166)		(0.277)
Age	0.032	-0.036	0.052	-0.044
	(0.118)	(0.134)	(0.111)	(0.140)
Education	0.124	0.125	0.120	0.121
	(0.118)	(0.067)	(0.120)	(0.067)
Work experience	-0.022	0.074	-0.028	0.081
	(0.115)	(0.098)	(0.113)	(0.097)
Observations	142	142	142	142
log likelihood	-76.695	-55.886	-77.020	-55.481

Note: Marginal effects from a probit model where the outcome is whether there is any punishment. Only run on observations with low price. “Dictator” is the punishment, measured in units of 100 NOK, that would implement the mean distribution chosen in the dictator game. Standard errors (in parentheses) corrected for clustering on individuals.