Essays on Mechanism Design

by

Douglas Scott Smith

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2011

Doctoral Committee:

Professor Tilman Börgers, Chair
Professor Yan Chen
Assistant Professor Stephen Lauermann
Associate Professor Lones Smith, University of Wisconsin

For my parents

ACKNOWLEDGEMENTS

I am deeply thankful to Tilman Börgers, my committee chair, whose encouragement, mentorship and patience have been crucial in the writing of this dissertation. He is also the coauthor of chapters 4 and 5.

I also wish to thank the other members of my committee, Yan Chen, Stephen Lauermann and Lones Smith, who all provided useful extremely useful feedback and advice.

Additionally, I am grateful to Mary Rigdon, my coauthor on chapter 5, for her guidance and collaboration.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

Mechanism design uses game theory to analyze how to construct mechanisms that give agents incentives to produce an outcome as optimal as possible for the mechanism designer. My dissertation is on robust mechanism design, where the mechanism designer has a great deal of uncertainty about what agents believe about other agents' preferences and beliefs. Much research on robust mechanism design has focused on finding "dominant strategy mechanisms," mechanisms where agents' equilibrium actions depend only on their preferences and not their beliefs about the other agents. Focusing on such mechanisms avoids the problem of uncertainty about agents' beliefs because it makes any assumptions about agents' beliefs irrelevant.

Chapters 2 through 5 of my dissertation explore how effective dominant strategy mechanisms are at achieving mechanism designers' goals. Mechanisms are compared by considering performance of the mechanisms for a wide range of possible agent preferences and beliefs, and then defining one mechanism as an improvement on another mechanism if it performs as well (and sometimes better) for every possible realization of agents preferences and beliefs. The first two chapters explore this question in the context of a public good problem. Chapter 2 demonstrates that there are mechanisms that improve on the dominant strategy mechanisms in the sense just described; Chapter 3 examines the usefulness of this particular efficiency concept by

providing an example of a mechanism that is both very simple and unimprovable. The fourth and fifth chapters are coauthored with Tilman Börgers and look at the same question for the voting problem, where dominant strategy mechanisms are (random) dictatorships. Chapter 4 shows that random dictatorship can be improved on if the mechanism designer wants to maximize agents expectation of their welfare, but not ex post welfare. Chapter 5 shows the same results when the designer wants to maximize the sum of two agents welfare.

The last chapter, Chapter 6, is coauthored with Mary Rigdon and reports on an experiment designed to investigate whether seemingly altruistic behavior is actually a response to strategic incentives, by using pre-commitment of some agents to test whether agents act differently when their actions cant influence others future actions. Our data support the hypothesis that agents are motivated by social concerns.

CHAPTER II

A Prior Free Efficiency Comparison for the Public Good Problem

2.1 Introduction

The public good problem with private information has long been considered an important problem in economics. The simplest version of this problem is as follows. A group of agents must decide whether to produce a single unit of an indivisible, non-excludable public good and how to pay for it. The cost of production is commonly known, but the value that each agent would gain from the production of the public good, although privately known to that agent, is not necessarily known by the other agents. Production of the good is efficient if the sum of the agents' valuations is larger than the production cost of the public good. But if the total value gained is less than the cost, then the public good should not be created. In this paper I focus on the case in which participation is voluntary: no agent can be forced to contribute to the production of the public good.

If the agents adopt a decision procedure in which first agents report their values for the public good, and the decision whether to produce the public good and the division of the cost among the agents depend on the reported values, then it is possible that some agents will try to free-ride by reporting lower valuations than they actually

have. They may expect that others' reported values may be sufficiently high that the public good is produced even with their lower report, but that their own contribution to the cost is lower than it would have been had they reported their values truthfully. Such strategic misrepresentation of private values may undermine the efficiency of the given procedure. Indeed, for some settings it can be shown that there is no way to avoid strategic misrepresentation as long as the decision procedure adopted by the agents ensures that the sum of agents' payments equals the production cost of the public good, and participation is voluntary.[1] It is therefore, in these settings, impossible to find a decision procedure that always arrives at efficient decisions.

This impossibility result inspires the study of a "second best problem:" This problem is typically conceptualized as that of a welfare-maximizing mechanism designer who must design a game (a mechanism) which determines a set of available actions for each agent to choose from. The mechanism then determines, based on the agent's choices, whether or not the public good is created, and which agents will pay what share of the cost. The mechanism designer predicts the agents' choices in any possible mechanism using a game-theoretic equilibrium concept, such as Bayesian equilibrium. If it is assumed that agents' beliefs are derived from a common prior on some type space, then the mechanism designer can be assumed to share the same common prior, and the mechanism designer can seek to maximize expected welfare where the expected welfare calculation is based on the common prior.

This problem has been studied in the literature where the literature has focused on relatively special type spaces. In particular, Güth and Hellwig (1986), Mailath and Postlewaite (1990), and Ledyard and Palfrey (2007) have solved this problem for the case that agents' types are independently distributed. The mechanisms that arise in these settings as second best do not have simple and intuitive interpretations.

[1]The Vickrey-Clarke-Groves mechanism (Vickrey (1961), Clarke (1971) and Groves (1973)) produces the efficient outcome, but is not budget balanced. d'Aspremont and Gerard-Varet (1979) show that efficient mechanisms that are budget balanced exist, but these violate voluntary participation as some agents may expect to be made worse off by participating.

However, they typically imply underproduction of the public good in comparison to the first best.

Starting with Wilson (1987), some economists have criticized the classic Bayesian approach to mechanism design on the grounds that it makes too strong assumptions regarding agents' beliefs about other agents' valuations, and about other agents' beliefs. These assumptions reflect the mechanism designer's perception of the environment for which he designs a mechanism. Critics of the classic approach argue that we should instead study second-best mechanisms for a mechanism designer who has a wider range of uncertainty about the agents' beliefs. The common prior assumption of classic Bayesian mechanism design may also be criticized as too strong. First, common priors rule out certain hierarchies of beliefs, such as the "agreeing to disagree" hierarchy. Second, even if agents' hierarchies of beliefs are such that they can be derived from a common prior, it is not clear why the mechanism designer should share this common prior, as classic Bayesian mechanism design implicitly assumes.

There are two responses to this concern with the Bayesian approach. One solution could be to study second best mechanisms with more complex type spaces, and allow the subjective belief of the mechanism designer to not necessarily be identical to the common prior that underlies agents' beliefs (if such a common prior exists). But with more complex priors the problem can become intractable. The more frequently chosen approach is to impose stronger solution concepts that make assumptions about agents' beliefs irrelevant. This approach means focusing on mechanisms where equilibrium strategies are (weakly) dominant or ex-post incentive compatible. When only such mechanisms are considered, assumptions about agents' beliefs are rendered innocuous. The question then becomes whether imposing these stronger solution concepts excludes too many mechanisms from consideration. If so, the approach will constrain the mechanism designer to a lower level of efficiency than might be achieved by directly analyzing the problem for a larger class of type spaces. Recent papers

(Bergemann and Morris, 2005; Chung and Ely, 2007) have explored this question for different mechanism design settings.

This paper proposes an alternative approach to the study of second best mechanisms for the public good problem. I assume there is some, possibly large, set of type spaces that the mechanism designer recognizes as possible descriptions of the agents' type space. Then one mechanism is said to improve on another mechanism if the former is at least as efficient as the latter for every realization of types on every type space, and strictly more efficient for at least one type space that the mechanism designer considers. When a mechanism designer considers a sufficiently rich set of type spaces, this is equivalent to making no assumption about the mechanism designer's perception of the agents' beliefs and private values. Furthermore, the induced partial ranking of mechanisms is independent of any beliefs the mechanism designer might have about the relative likelihood of different type spaces and realizations of types within each type space. In that sense the ranking is robust. Furthermore, I prove a result regarding the soundness of the improvability ranking by showing that, if the typespaces the mechanism designer considers are finite, any feasible mechanism is either unimprovable or there is an unimprovable mechanism that improves upon it.

I use the improvability ranking to examine the effect of restricting attention to mechanisms where agents' equilibrium actions depend only on their private values and not on their beliefs. In private value settings those mechanisms are dominant strategy mechanisms. I first characterize the set of dominant strategy mechanisms in this setting (I restrict attention to deterministic mechanisms). I then show that in this setting every dominant strategy mechanism is improved on by some mechanism, i.e., there is some mechanism that is at least as efficient as the dominant strategy mechanism on any type space, and strictly more efficient on some type space. While this mechanism improves on the dominant strategy mechanism from an efficiency perspective, it is just as "detail free." Furthermore, any sufficiently rich set of possible

type spaces would lead a welfare maximizer to prefer these non-dominant strategy mechanisms to the class of dominant strategy mechanisms. This analysis implies that in the public good setting, restricting attention to dominant strategy mechanisms is not a satisfactory solution to finding optimal mechanisms on richer type spaces.

A further goal of this research is to characterize the unimprovable mechanisms in the public good problem. I explore this question in a companion paper, "An Unimprovability Result for the Public Good Problem." In that paper, I show a simple mechanism that is unimprovable among finite-action mechanisms on the universal type space.

There is a natural connection between the literature on public good mechanisms without hidden information and the mechanisms I find that improve on the dominant strategy mechanisms. The improving mechanisms allow for agents to reduce the cost of the public good to other agents (and pay the difference in cost themselves). In this way they are similar in spirit to the public good "compensation mechanisms" studied by Varian (1994a and 1994b), where agents contributing to a public good (in a complete information setting) can subsidize other agents' purchases of the public good, and in so doing make those agents internalize the social benefit of their purchase of the public good. In a similar fashion, the mechanism that improves on a dominant strategy mechanism allows an agent to express a stronger preference for a public good by reducing what another agent has to pay for the good to be produced.

This paper contributes to the literature examining when a mechanism designer might choose to use a mechanism with belief-invariant strategies. One important difference between this paper and the previous literature is the improvability comparison. Bergemann and Morris (2005) look at a general mechanism design setting and a mechanism designer who wants to implement a social choice rule. They prove results describing environments in which a mechanism designer can restrict attention to mechanisms with belief-invariant equilibria and not reduce the set of social choice

rules that can be implemented. In their analysis, social choice rules only depend on payoff-relevant information. In contrast, the improvability concept of this paper takes a description of the mechanism designer's preference over outcomes that depends only on agents' valuations and then creates a ranking of mechanisms that reflects the goal of finding, given that objective, a second best mechanism. This ranking can make a distinction between mechanisms when for a given profile of pay-off relevant types, one mechanism implements the optimal outcome for more possible beliefs of the agents than another mechanism.

Chung and Ely (2007) examine an auction setting where the mechanism designer is trying to maximize revenue, has a prior over the distribution of agents' valuations, and evaluates mechanisms on the basis of their worst-case revenue outcome over all possible beliefs of the agents (a *maxmin* approach). They show that it in this setting the mechanism designer can rationally choose a dominant strategy mechanism. A difference between Chung and Ely's setting and the setting in this paper is that they assume that the mechanism designer only cares about the worst-case outcome, whereas in this paper the mechanism designer compares outcomes across all possible agent realizations. Given two mechanisms that have the same worse-case outcome, one mechanism can still improve on the other mechanism if the former does strictly better than the latter on other type spaces. However, maxmin preferences would not make a distinction between the performance of the two mechanisms.

Chung and Ely also show that for a dominant strategy mechanism that is optimal among dominant strategy mechanisms for a given distribution over agents' valuations, they can find a type space that describes agents' beliefs such that a dominant strategy mechanism is a rationalizable choice on that type space. In contrast, this paper's results suggest in the public goods setting, if the mechanism designer takes into consideration a sufficiently broad range of types spaces, any dominant strategy mechanism will be strictly inferior from an expected welfare standpoint to some

mechanism where equilibrium actions depend on beliefs as well as payoff types.

An interesting recent paper by Yamashita (2011), written independently from and at the same time as this paper, looks at a related question in the bilateral trade setting. The approach of the paper differs from this paper in that in that it constructs a robust mechanism design approach by analyzing outcomes that can arise from a given mechanism when agents play non-weakly dominated strategies, and evaluates mechanisms based on the minimal welfare the mechanism can guarantee when agents can potentially play any non-weakly dominated strategy. He shows that for some assumptions on the distribution of agents' valuations, dominant strategy mechanisms (specifically, fixed price mechanisms) can be optimal. He also shows that under certain assumptions on the agent's type spaces, a two-price mechanism that is not a dominant strategy mechanism not only performs better than any dominant strategy mechanism, but is in fact optimal among the class of finite mechanisms in terms of its minimal welfare guarantee.

The paper is structured as follows: in the next section, I introduce the public good problem and define the set of type spaces that a mechanism designer may consider. In section 2.3, I describe the set of feasible mechanisms and equilibria. Section 2.4 defines the concept of improvability as a (partial) ranking of mechanisms. Section 2.5 defines dominant strategy mechanisms, while section 2.6 shows that in thinking about the efficiency of dominant strategy mechanisms we can focus on a particular set of maximally efficient dominant strategy mechanisms, the fixed contribution mechanisms. In section 2.7 I describe another mechanism, the "additional contribution mechanism." Section 2.8 compares the efficiency of dominant strategy mechanisms and interim implementable mechanisms by showing that every fixed contribution mechanism, and hence any dominant strategy mechanism, is improved on by some additional contribution mechanism for any mechanism designer who considers a sufficiently rich set of type spaces. Section 2.9 concludes.

2.2 The Public Good Problem

A set of N individuals, indexed by $i \in \{1, 2, ...N\}$, have to decide whether an indivisible and non-excludable public good will be created, and how it will be paid for. Each individual derives a private benefit from the creation of the public good, and each individual knows their own benefit (valuation) but not necessarily the valuations of the other individuals. Preferences are assumed to have a quasi-linear structure, and monetary transfers between agents are possible. In this paper I focus on the problem of a welfare maximizing social planner, who faces a budget balance constraint as well as individual rationality constraints, who wants the public good to be created if and only if the sum of the private benefits is at least as large as the cost of producing the public good. The rest of this section develops the formal model and notation.

Let the set of outcomes be $Y = \{0, 1\} \times \mathbf{R}^N$ where the first component is interpreted as the probability that the public good is created, and the other components are the transfers from (or if negative, to) each individual. Let $y \equiv (q, \tau) \in Y$ where $\tau = (\tau_1, \tau_2, ..., \tau_N)$. The cost of creating the public good is c.

Individual i's valuation of the good is $v_i \in [\underline{v}, \overline{v}]$, with the valuation profile $v = (v_1, v_2, ..., v_N)$. I assume that $N\overline{v} > c > N\underline{v}$, so if all agents have the highest valuation then the value created by the public is greater than its costs, and if all agents have the lowest valuation then the value created by the public good is less than its cost.

I assume that each agent's utility only depends on the outcome and their valuation of the good. Furthermore, utilities are linear in the probability of the public good being created and the monetary transfer:

$$u_i(y, v_i) = v_i \cdot q - \tau_i \tag{2.1}$$

Define a type space $\mathbf{T} \equiv (T_i, \hat{v}_i, \hat{\pi}_i)_{i \in \{1,...,N\}}$ as a collection of types, T_i, for each

player ($t_i \in T_i$), a function for each agent i from types to valuations of the good, $\hat{v}_i : T_i \to [\underline{v}, \overline{v}]$, and a function for each agent i that maps from agent i's types to beliefs over other agents' types, $\hat{\pi}_i : T_i \to \Delta(T_{-i})$, where $\Delta(T_{-i})$ is the set of probability distributions over T_{-i}. I will write $\hat{\pi}_{t_i}[E]$ for the probability that agent i of type t_i puts on the other agents being of a type profile in the set E. Furthermore, define $T = \times T_i$, and for any i, $T_{-i} = \times_{j \neq i} T_j$.

A given typespace T can be any set of types (and valuation and beliefs functions) as long as a suitable measure can be defined for beliefs on T_{-i} for all i. I do not require $\mathbf{T}$ to be a subset of the universal type space (see Mertens and Zamir (1985) for a description of the universal type space). In particular, I allow $\mathbf{T}$ to contain "redundant" types, that is types with the same beliefs and valuations. I will focus on two main cases of type spaces. I'll call $\mathbf{T}$ *finite* if $|T_i|$ is finite for all i. I will also consider the universal type space.

I model the mechanism designer as having in mind a set $\mathbf{\Omega}$ of type spaces which the mechanism designer believes contains the actual type space. The mechanism designer then considers mechanisms and their equilibria on all the type spaces in $\mathbf{\Omega}$.

2.3 Feasible Mechanisms

A mechanism $M = (A, \hat{y})$ consists of a set of actions A_i for each agent, with $A = \times A_i$, and a mapping from combinations of actions into outcomes, $\hat{y} : A \to Y$. In particular, I define $\hat{y}(a) \equiv (q(a), \tau(a))$. Sequential games are included in the analysis through their strategic form representations. A (Bayesian Nash) equilibrium of a mechanism is defined in the following way:

Definition II.1. An equilibrium a^* of a mechanism M consists of a mapping for each type space $\mathbf{T} \in \mathbf{\Omega}$, and for each agent, $a^*_{\mathbf{T},i} : T_i \to A_i$, such that for all i and all $t_i \in T_i$,

$a^*_{\mathbf{T},i}(t_i)$ is a best response to the other agent's equilibrium actions $a^*_{\mathbf{T},-i} = \times_{j \neq i}\ a^*_{\mathbf{T},j}$, i.e.

$$E_{\hat{\pi}(t_i)}\left[\hat{v}_i(t_i) \cdot q\left(a^*_{\mathbf{T},i}(t_i), a^*_{\mathbf{T},-i}(t_{-i})\right) - \tau_i\left(a^*_{\mathbf{T},i}(t_i), a^*_{\mathbf{T},-i}(t_{-i})\right)\right]$$

$$\geq E_{\hat{\pi}(t_i)}\left[\hat{v}_i(t_i) \cdot q\left(\hat{a}_i, a^*_{\mathbf{T},-i}(t_{-i})\right) - \tau_i\left(\hat{a}_i, a^*_{\mathbf{T},-i}(t_{-i})\right)\right] \tag{2.2}$$

for all $\hat{a}_i \in A_i$.

Note that an equilibrium defines actions for all agents and for all type spaces $\mathbf{T} \in \mathbf{\Omega}$.

Given an equilibrium a^*, I can define for any $\mathbf{T}$ the expected utility of a type $t_i \in T_i$:

$$U_i(t_i) \equiv E_{\hat{\pi}(t_i)}\left[\hat{v}_i(t_i) \cdot q(a^*_{\mathbf{T},i}(t_i), a^*_{\mathbf{T},-i}(t_{-i})) - \tau_i(a^*_{\mathbf{T},i}(t_i), a^*_{\mathbf{T},-i}(t_{-i}))\right]$$

and the expected utility of t_i playing action a_i in equilibrium,

$$U_i(t_i, a_i) \equiv E_{\hat{\pi}(t_i)}\left[\hat{v}_i(t_i) \cdot q(a_i, a^*_{\mathbf{T},-i}(t_{-i})) - \tau_i(a_i, a^*_{\mathbf{T},-i}(t_{-i}))\right].$$

Similarly, define the expected likelihood of the public good being produced for t_i in equilibrium,

$$Q_i(t_i) \equiv E_{\hat{\pi}(t_i)} q(a^*_{\mathbf{T},i}(t_i), a^*_{\mathbf{T},-i}(t_{-i}))$$

and the expected likelihood of the public good being produced for t_i in equilibrium if t_i plays a_i,

$$Q_i(t_i, a_i) \equiv E_{\hat{\pi}(t_i)} q(a_i, a^*_{\mathbf{T},-i}(t_{-i})).$$

Definition II.2. $M = (A, \hat{y})$ and an equilibrium a^* of that mechanism are **feasible** if the following two conditions are satisfied:

- Ex-post Budget Balance (BB). For all $a \in A$,

$$\sum_{i=1}^{N} \tau_i(a) = q(a) \cdot c \tag{2.3}$$

Note that budget balance is a condition on the mechanism, not on the equilibrium.

- Interim individual rationality (IIR). For all $\mathbf{T} \in \mathbf{\Omega}$ and all $t_i \in T_i$,

$$U_i(t_i) \geq 0 \tag{2.4}$$

The definition of equilibrium implies a third property:

- Bayesian Incentive Compatibility (BIC). For all $\mathbf{T} \in \mathbf{\Omega}$ and all $t_i \in T_i$,

$$U_i(t_i) \geq E_{\hat{\pi}(t_i)} \left[\hat{v}_i(t_i) \cdot q(\hat{a}_i, a^*_{\mathbf{T},-i}(t_{-i})) - \tau_i(\hat{a}_i, a^*_{\mathbf{T},-i}(t_{-i})) \right] \tag{2.5}$$

$\forall \hat{a}_i \in A_i$

I now turn to comparisons of feasible mechanisms.

2.4 Improvability

The mechanism designer is assumed to want to maximize the sum of agents' utilities. I further assume the mechanism designer must choose a feasible mechanism. Therefore budget balance implies that the mechanism designer's value function at any realization of types t and for any outcome $y = (q, \tau)$ is equal to,

$$V(y, t) = q \cdot \left(\sum_{i=1}^{N} \hat{v}_i(t_i) - c \right) \tag{2.6}$$

for any $\mathbf{T} \in \mathbf{\Omega}$ and $t \in T$.

I make no assumptions about the mechanism designer's beliefs. This allows the analysis to apply to a mechanism designer with any Bayesian beliefs who maximizes expected welfare, but can accomodate a mechanism designer with different or less well defined beliefs, or different preferences. Instead I develop a prior-free comparison of the efficiency of two mechanisms. I say a feasible mechanism M and equilibrium a^* improve on another mechanism $\tilde{M}$ and equilibrium $\tilde{a}^*$ if a^* produces at least as efficient a result for any realization of types as $\tilde{a}^*$, and for some possible realization of types produces a strictly more efficient result. Efficiency is judged in terms of equation (2.6).

Definition II.3. A feasible mechanism M and equilibrium a^* **improve on** a feasible mechanism $\tilde{M}$ and equilibrium $\tilde{a}^*$ if for all type spaces $\mathbf{T} \in \mathbf{\Omega}$ and all $t \in T$,

$$\begin{aligned} q^M(a^*_{\mathbf{T}}(t)) \geq q^{\tilde{M}}(\tilde{a}^*_{\mathbf{T}}(t)) \quad & \text{when } \sum_{i=1}^{N} \hat{v}_i(t_i) > c \\ & \text{and} \\ q^M(a^*_{\mathbf{T}}(t)) \leq q^{\tilde{M}}(\tilde{a}^*_{\mathbf{T}}(t)) \quad & \text{when } \sum_{i=1}^{N} \hat{v}_i(t_i) < c \end{aligned} \tag{2.7}$$

and there is some $\mathbf{T} \in \mathbf{\Omega}$ and $t \in T$ such that the applicable inequality in (2.7) is strict.

This comparison corresponds to a weak dominance approach to comparing mechanisms (and associated equilibria). It defines a partial ranking of mechanisms that will partially describe a mechanism designer's preferences under a broad range of assumptions about the mechanism designer. Specifically, the analysis will apply for any mechanism designer who prefers a mechanism over another if the former is always at least as efficient and sometimes strictly more efficient than the latter. There are mod-

els of mechanism designers that would not fit this description (for example, Chung and Ely (2007)'s *maxmin* preferences disregard performance in all but the worst-case situation). Even in these cases, however, the intuitive appeal of the improvability criterion make it an appealing potential tie-breaker between mechanisms that are ranked equally according to the specified preferences.

This definition of improvability focuses on ex-post outcomes rather than an interim measure of efficiency. That choice reflects the description of the mechanism designer's preference for maximizing realized aggregate welfare. One reason for using realized outcomes is that the mechanism designer may have different beliefs about the distribution of types than any particular agent, and so may disagree with an agent's interim evaluation of a mechanism. (The mechanism designer's beliefs are not formally modeled in this paper.) Furthermore, if agents have inconsistent beliefs, it may be interim improving to allow agents to make bets between themselves. Focusing on ex-post outcomes avoids entangling the analysis of the public goods problem with the welfare analysis of bets reflecting inconsistent beliefs. Although any given mechanism may allow agents to make such bets, the evaluation of the mechanism depends only on agents' realized utilities.

The following straightforward lemma establishes that improvability relates in a natural way to ex-post welfare maximization.

Lemma II.4. *A feasible mechanism $M = (A^M, \hat{y}^M)$ and equilibrium a^* improve on a feasible mechanism $\tilde{M} = (A^{\hat{M}}, \hat{y}^{\hat{M}})$ and equilibrium $\tilde{a}^*$ if for all type spaces $\mathbf{T} \in \mathbf{\Omega}$ and all $t \in T$,*

$$\sum_{i=1}^{N} u_i \left(\hat{y}^M(a^*_{\mathbf{T}}(t)), \hat{v}_i(t_i) \right) \geq \sum_{i=1}^{N} u_i \left(\hat{y}^{\tilde{M}}(\tilde{a}^*_{\mathbf{T}}(t)), \hat{v}_i(t_i) \right) \tag{2.8}$$

and for some $\mathbf{T} \in \mathbf{\Omega}$ *and* $t \in T$ *the inequality is strict.*

Proof. Observe that

$$\sum_{i=1}^{N} u_i \left(\hat{y}^M(a^*_{\mathbf{T}}(t)), \hat{v}_i(t_i) \right)$$

$$= q^M \left(a^*_{\mathbf{T}}(t) \right) \left(\sum_{i=1}^{N} \hat{v}_i(t_i) - c \right) - \left(\sum_{i=1}^{N} \tau_i^M(a^*_{\mathbf{T}}(t)) - c \cdot q^M \left(a^*_{\mathbf{T}}(t) \right) \right),$$

which, by budget balance,

$$= q^M \left(a^*_{\mathbf{T}}(t) \right) \left(\sum_{i=1}^{N} \hat{v}_i(t_i) - c \right). \tag{2.9}$$

Therefore the sum of realized utilities will be

$$\begin{aligned} &\text{increasing in } q^M \left(a^*_{\mathbf{T}}(t) \right) && \text{when } \sum_{i=1}^{N} \hat{v}_i(t_i) > c \\ &\text{and decreasing in } q^M \left(a^*_{\mathbf{T}}(t) \right) && \text{when } \sum_{i=1}^{N} \hat{v}_i(t_i) < c. \end{aligned}$$

So M and a^* improve on another mechanism $\tilde{M}$ and $\tilde{a}^*$ if M and a^* lead to a (weakly) greater sum of realized utilities for every t, and a strictly greater sum of realized utilities for some t.

□

The relationship between improvability and the sum of realized utilities holds because of strict budget balance. With a non-negative budget balance the relationship between the two becomes more complicated, although the comparison of dominant strategy mechanisms and Bayesian implementable mechanisms still holds. See appendix B for details.

To confirm the soundness of improvability as a method of (partially) ranking mechanisms, it would be useful to know that for every improvable mechanism, there

exists an unimprovable mechanism that improves on it. The following result shows that this in fact the case in the finite type spaces setting. (I conjecture that a similar result holds for infinite type spaces.)

Proposition II.5. *Suppose* $\mathbf{\Omega}$ *is a set of finite type spaces. For any mechanism* M *and equilibrium* a^**, either* M *an* a^* *are unimprovable on* $\mathbf{\Omega}$ *or there exists an unimprovable mechanism* $\hat{M}$ *and equilibrium* $\hat{a}^*$ *that improve on* M *and* a^* *on* $\mathbf{\Omega}$*.*

Proof. Suppose M and a^* are improvable. For any $\mathbf{T}$, we can define the following restricted improvability concept:

Definition II.6. A feasible mechanism M and equilibrium a^* **improve on** a feasible mechanism $\tilde{M}$ and equilibrium $\tilde{a}^*$ **on** $\mathbf{T}$ if for all $t \in T$,

$$\begin{aligned} q^M(a^*_{\mathbf{T}}(t)) \geq q^{\tilde{M}}(\tilde{a}^*_{\mathbf{T}}(t)) \quad & \text{when } \sum_{i=1}^{N} \hat{v}_i(t_i) > c \\ & \text{and} \\ q^M(a^*_{\mathbf{T}}(t)) \leq q^{\tilde{M}}(\tilde{a}^*_{\mathbf{T}}(t)) \quad & \text{when } \sum_{i=1}^{N} \hat{v}_i(t_i) < c \end{aligned} \tag{2.10}$$

and for some $t \in T$ such the applicable inequality in (2.10) is strict.

This is just the improvability definition with $\mathbf{\Omega} = \{\mathbf{T}\}$.The following result will be important in proving the proposition.

Lemma II.7. *For any feasible mechanism* M *and equilibrium* a^* *on finite* $\mathbf{T}$*, either* M *and* a^* *are unimprovable on* $\mathbf{T}$ *or there exists an unimprovable mechanism* $\hat{M}$ *and equilibrium* $\hat{a}^*$ *that improve on* M *and* a^* *on* $\mathbf{T}$*.*

Proof. If M and a^* are improvable then there must be some mechanism M_1 and equilibrium a^{*1} that improve on them. M_1 and equilibrium a^{*1} must be efficient on

strictly more type profiles than M and a^*. If M_1 and equilibrium a^{*1} are improvable, then there exists some M_2 and equilibrium a^{*2} that must be efficient on strictly more type profiles than M_1 and equilibrium a^{*1}. Repeating this process, we can create a sequence of mechanisms such that M_{k+1} and equilibrium a^{*k+1} must be efficient on strictly more type profiles than M_k and equilibrium a^{*k} for all $k \geq 1$. There are $\Pi_{i=1}^{N}|T_i|$ type profiles in $\mathbf{T}$, so the number of type profiles is finite. Therefore there exists some $k' \leq \Pi_{i=1}^{N}|T_i|$ such that $M_{k'}$ and equilibrium $a^{*k'}$ are unimprovable. $\square$

Let $\mathbf{\Omega}$ be a set of finite type spaces (possibly the set of all finite type spaces). For each $\mathbf{T} \in \mathbf{\Omega}$ select a $M_{\mathbf{T}}$ and equilibrium $a^{*\mathbf{T}}$ such that $M_{\mathbf{T}}$ and $a^{*\mathbf{T}}$ are unimprovable on $\mathbf{T}$ and improve on M and a^* on $\mathbf{T}$, or are equivalent to M and a^* if M and a^* are unimprovable on $\mathbf{T}$. Define the mechanism $M_{\mathbf{\Omega}}$ and equilibrium $a^{*\mathbf{\Omega}}$ by the following:

For each $\mathbf{T} \in \mathbf{\Omega}$, $A_i^{\mathbf{T}} = \{\mathbf{T}\} \times A_i^{M_{\mathbf{T}}}$ for all i.

Then $A_i^{M_{\mathbf{\Omega}}} = \cup_{\mathbf{T} \in \mathbf{\Omega}} A_i^{\mathbf{T}}$ for all i.

Define $a_i \in A_i^{M_{\mathbf{\Omega}}} = (\mathbf{T}_i, a_i^{\mathbf{T}})$, and $a^{\mathbf{T}} = \left(a_1^{\mathbf{T}}, ..., a_N^{\mathbf{T}}\right)$. Then let

$$\hat{y}^{M_{\mathbf{\Omega}}}(a) = \begin{cases} \hat{y}^{M_{\mathbf{T}}}(a^{\mathbf{T}}) & \text{if } \mathbf{T}_i = \mathbf{T}_j \ \forall i, j \\ \\ (0, \vec{0}) & \text{otherwise.} \end{cases}$$

For any $\mathbf{T}'$ and any $t_i \in T_i'$, let $a_{\mathbf{T}',i}^{*\mathbf{\Omega}}(t_i) = \left(\mathbf{T}', a_{\mathbf{T}',i}^{*\mathbf{T}'}(t_i)\right)$.

An agent's action is an indication of a type space $\mathbf{T}$ and an action from those actions available in $M_{\mathbf{T}}$. By inspection $a^{*\mathbf{\Omega}}$ is an equilibrium of $M_{\mathbf{\Omega}}$. (An agent has no incen-

tive to indicate a different type space from the other agents, and given agreement on the type space the agent has no incentive deviate from the equilibrium $a^{*\mathbf{T}}$). Furthermore, for any $\mathbf{T}$, $M_{\mathbf{\Omega}}$ and $a^{*\mathbf{\Omega}}$ are unimprovable on $\mathbf{T}$ because $M_{\mathbf{T}}$ and equilibrium $a^{*\mathbf{T}}$ are unimprovable on $\mathbf{T}$ and for all $t \in T$ we have $\hat{y}^{M_{\mathbf{\Omega}}}(a^{*\mathbf{\Omega}}_{\mathbf{T}}(t)) = \hat{y}^{M_{\mathbf{T}}}(a^{*\mathbf{T}}_{\mathbf{T}}(t))$.

That $M_{\mathbf{\Omega}}$ and $a^{*\mathbf{\Omega}}$ are unimprovable on $\mathbf{T}$ for all $\mathbf{T} \in \mathbf{\Omega}$ implies they are unimprovable, by the following logic: if any mechanism $\hat{M}$ and equilibrium $\hat{a}^*$ improved on $M_{\mathbf{\Omega}}$ and $a^{*\mathbf{\Omega}}$, that would imply there exists a $\mathbf{T}'$ such that $\hat{M}$ and $\hat{a}^*$ improved on $M_{\mathbf{\Omega}}$ and $a^{*\mathbf{\Omega}}$ on $\mathbf{T}'$, which would be a contradiction.

Furthermore, M and a^* improvable implies that M and a^* are improvable on some $\mathbf{T}'$, and by construction $M_{\mathbf{\Omega}}$ and $a^{*\mathbf{\Omega}}$ improve on M and a^* on $\mathbf{T}'$. □

2.5 Dominant Strategy Mechanisms

In this section I describe mechanisms that have equilibria that are *belief-invariant*, that is agents' equilibrium actions depend only on their valuations and not on their beliefs. In the following section I describe the fixed contribution mechanism, and establish that every belief-invariant equilibrium of an interim individually rational and budget balanced mechanism is either equivalent to a fixed contribution mechanism or improved on by a fixed contribution mechanism.

For ease of exposition I assume in this section that $\mathbf{\Omega}$ includes all finite type spaces, or the universal type space, or both. This allows me to assume that agents with any valuation profile (and with certain beliefs) exist in some $\mathbf{T}$ in $\mathbf{\Omega}$. This is consistent with the usual motivation for using dominant strategy mechanisms, that they are robust to the specification of the type space.

Definition II.8. An equilibrium a^* of a mechanism $M = (A^M, \hat{y}^M)$ is **belief-**

invariant if for each i there exists a function $\hat{a}_i : [\underline{v}, \overline{v}] \to A_i^M$ such that for all $\mathbf{T} \in \mathbf{\Omega}$ and all $t_i \in T$, $a^*_{\mathbf{T},\mathbf{i}}(t_i) = \hat{a}_i(\hat{v}_i(t_i))$.

If we restrict attention to equilibria where actions only depend on valuations, then the following definitions are natural:

Definition II.9. A mechanism $M = (A^M, \hat{y}^M)$ is a **direct mechanism** if $A_i^M = [\underline{v}, \overline{v}]$ for all i.

Definition II.10. A belief-invariant equilibrium a^* (and the associated $\hat{a}_i$) of a direct mechanism is a **truth-telling equilibrium** if for all i and all $v_i \in [\underline{v}, \overline{v}]$, $\hat{a}_i(v_i) = v_i$.

Given these definitions, I can apply the revelation principle:

Lemma II.11. *Given a belief-invariant equilibrium a^* (and the associated $\hat{a}$) of a mechanism $\hat{M} = (A^{\hat{M}}, \hat{y}^{\hat{M}})$, the direct mechanism $M = (A^M, \hat{y}^M)$ defined by $\hat{y}^M(v) = \hat{y}^{\hat{M}}(\hat{a}(v))$ for all $v \in [\underline{v}, \overline{v}]^N$ has a truth-telling equilibrium $\overline{a}^*$, defined by*

$$\overline{a}^*_{\mathbf{T},\mathbf{i}}(t_i) = \hat{v}_i(t_i) \tag{2.11}$$

for all $\mathbf{T} \in \mathbf{\Omega}$ and all t_i, where

$$\hat{y}^M(\overline{a}^*_{\mathbf{T}}(t)) = \hat{y}^{\hat{M}}(a^*_{\mathbf{T}}(t)) \tag{2.12}$$

for all $\mathbf{T} \in \mathbf{\Omega}$ and all t.

Proof. By construction, $\hat{y}^M(\overline{a}^*_{\mathbf{T}}(t)) = \hat{y}^M(\hat{a}(\hat{v}(t)) = \hat{y}^{\hat{M}}(a^*_{\mathbf{T}}(t))$. I show $\overline{a}^*$ is an equilibrium of M by contradiction. If $\overline{a}^*$ is not an equilibrium, then there exists a $\mathbf{T} \in \mathbf{\Omega}$ and t_i such that an agent with type t_i and who takes action $\hat{v}_i(t_i)$ would want to deviate to another action v' in $A_i^M = [\underline{v}, \overline{v}]$. Because t_i deviating to v' has the same effect on outcomes as t_i deviating from $\hat{a}_i(\hat{v}_i(t_i))$ to $\hat{a}_i(v')$ when agents are in the a^*

equilibrium of $\hat{M}$, it must be that t_i has an incentive to deviate under a^* as well, which contradicts the assumption a^* is an equilibrium.

Let $\tilde{a}^* : [\underline{v}, \overline{v}] \to [\underline{v}, \overline{v}]$ be the identity function. Then for all $\mathbf{T}$ and all $t_i \in T_i$, $\overline{a}_i^*(t_i) = \tilde{a}^*(\hat{v}_i(t_i))$ which shows $\overline{a}^*$ is a truth-telling equilibrium.

□

Definition II.12. A direct mechanism satisfies **dominant strategy incentive compatibility** (DIC) if, for all v,

$$v_i \cdot q(v) - \tau_i(v) \geq v_i \cdot q(\hat{v}, v_{-i}) - \tau_i(\hat{v}, v_{-i}) \qquad \forall \hat{v} \in [\underline{v}, \overline{v}] \tag{2.13}$$

A direct mechanism is called a **dominant strategy mechanism** if it satisfies dominant strategy incentive compatibility.

Lemma II.13. *A direct mechanism M has a truth-telling equilibrium a^* if and only if it is a dominant strategy mechanism.*

Proof. If M has a truth-telling equilibrium a^*, and $\mathbf{\Omega}$ contains all finite type spaces or the universal type space, then for any $v_i \in [\underline{v}, \overline{v}]$ and $v_{-i} \in [\underline{v}, \overline{v}]^{N-1}$, we can find a $\mathbf{T}$ where there exists a type t_i with $\hat{\pi}_{t_i}\{t_{-i} | \hat{v}_{-i}(t_{-i}) = v_{-i}\} = 1$. Then a^* being a truth-telling equilibrium implies that for all v_i',

$$E_{\hat{\pi}_{t_i}} u_i(t_i, \hat{v}_i(t_i)) = v_i \cdot q^M(v) - \tau_i^M(v)$$

$$\geq E_{\hat{\pi}_{t_i}} u_i(t_i, v_i') = v_i \cdot q^M(v_i', v_{-i}) - \tau_i^M(v_i', v_{-i}).$$

Therefore M satisfies dominant strategy incentive compatibility.

If a direct mechanism is a dominant strategy mechanism, then it is immediate that $\hat{a}_i(v_i) = v_i$ is truth-telling equilibrium.

□

Lemma II.14. *A dominant strategy mechanism and its truth telling equilibrium are feasible if and only if they satisfy the following two properties:*

- Ex-post Budget Balance (BB). For every $v \in [\underline{v}, \overline{v}]^N$, it must be that $\sum_{i=1}^{N} \tau_i(v) = q(v) \cdot c$.

- Ex-post individual rationality (EIR). For all i, all v,

$$v_i \cdot q(v) - \tau_i(v) \geq 0 \tag{2.14}$$

Proof. This characterization of ex-post budget balance is immediate from the ex-post budget balance property of feasible mechanisms. Therefore it is immediate that if a dominant strategy mechanism and its truth telling equilibrium are feasible they satisfy ex-post budget balance, and if they don't satisfy it then they cannot be feasible.

For ex-post individual rationality, just as in the proof of Lemma II.13, for any $v_i \in [\underline{v}, \overline{v}]$ and $v_{-i} \in [\underline{v}, \overline{v}]^{N-1}$, it is possible to construct a $\mathbf{T}$ such that there exists a type t_i with $\hat{\pi}_{t_i}\{t_{-i} | \hat{v}_{-i}(t_{-i}) = v_{-i}\} = 1$ (assuming that $\mathbf{\Omega}$ includes all finite type spaces or the universal type space). Then

$$E_{\hat{\pi}_{t_i}} u_i(\hat{y}(v), \hat{v}_i(t_i)) = v_i \cdot q(v) - \tau_i(v)$$

and interim individual rationality therefore implies $v_i \cdot q(v) - \tau_i(v) \geq 0$.

If a dominant strategy mechanism and its truth telling equilibrium satisfy EIR, then for all i and all v, $q(v) \cdot v_i - \tau_i(v) \geq 0$, so

$$U_i(t_i) = E_{\hat{\pi}(t_i)}\left[q(v) \cdot v_i - \tau_i(v)\right] \geq 0 \tag{2.15}$$

□

2.6 Fixed Contribution Mechanisms

Now I turn to a description of a particular class of mechanisms that are dominant strategy mechanisms, the fixed contribution mechanisms. First I define the indirect mechanism, then describe an equilibrium, which allows me to define the direct mechanism version of the fixed contribution mechanism.

The fixed contribution mechanism is defined as follows: the mechanism designer chooses a set of cost shares $s = (s_1, ..., s_N)$ such that $\sum_{i=1}^{N} s_i = c$.

Each agent i learns his/her type, and is asked whether he/she wishes to have the public good created if they have to pay s_i, and reply "Yes" or "No." If all agents indicate "Yes," the public good is created, with each agent paying s_i. Otherwise the public good is not created, and no transfers are made.

Formally, $A_i^F = \{\text{“}Yes\text{”}, \text{“}No\text{”}\}$, and $\hat{y}^F(a)$ is defined as follows:

$$q^F(a) = \begin{cases} 1 & \text{if } a_i = \text{“Yes” } \forall i \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

$$\tau^F(a) = \begin{cases} \text{s} & \text{if } a_i = \text{“Yes” } \forall i \\ 0 & \text{otherwise} \end{cases} \tag{2.17}$$

To consider strategies of the agents, fix a type space $\mathbf{T}$. A pure strategy for an agent is a mapping $a_i : T_i \to \{\text{“Yes”}, \text{“No”}\}$ that indicates for each type of the agent whether they agree to the public good being created or not.

Define the following strategy for each agent i:

$$a_i^*(t_i) = \begin{cases} \text{“Yes”} & \text{if } \hat{v}_i(t_i) \geq s_i \\ \text{“No”} & \text{if } \hat{v}_i(t_i) < s_i \end{cases} \tag{2.18}$$

The following lemma shows the straightforward result that these strategies form an equilibrium (and are in fact weakly dominant for all i whenever $\hat{v}_i(t_i) \neq s_i$).

Lemma II.15. *The strategies a_i^* for all i form a belief-invariant Bayesian Nash equilibrium on any type space $\mathbf{T}$. Furthermore, agents play weakly dominant strategies when (for any i) their valuation is such that $\hat{v}_i(t_i) \neq s_i$.*

Proof. We show that each agent's strategy is a weakly dominant strategy.

If the agent's valuation $\hat{v}_i(t_i) > s_i$,

$$u_i\left(y(\text{“Yes”}, a_{-i}(t_{-i})), \hat{v}_i(t_i)\right) \in \{0, \hat{v}_i(t_i) - s_i\}, \quad \forall t_{-i} \in T_{-i}$$

Therefore, regardless of which value it takes on,

$$u_i\left(y(\text{“Yes”}, a_{-i}(t_{-i})), \hat{v}_i(t_i)\right) \geq 0 = u_i\left(y(\text{“No”}, a_{-i}(t_{-i})), \hat{v}_i(t_i)\right).$$

and if $a_j(t_j) =$“Yes” for all $j \neq i$, then the inequality is strict. Therefore “Yes” is a weakly dominant.

If the agent's valuation $\hat{v}_i(t_i) < s_i$,

$$u_i\left(y(\text{“Yes”}, a_{-i}(t_{-i})), \hat{v}_i(t_i)\right) \in \{0, \hat{v}_i(t_i) - s_i\}, \quad \forall t_{-i} \in T_{-i}$$

Therefore, regardless of which value it takes on,

$$u_i\left(y(\text{“Yes”}, a_{-i}(t_{-i})), \hat{v}_i(t_i)\right) \leq 0 = u_i\left(y(\text{“No”}, a_{-i}(t_{-i})), \hat{v}_i(t_i)\right).$$

and if $a_j(t_j) =$ “Yes” for all $j \neq i$, then the inequality is strict. Therefore “No” is a

weakly dominant strategy.

Since each agent is playing a weakly dominant strategy, or is indifferent between either action, the strategies a_i^* for all i form a Bayesian Nash equilibrium. □

I call this equilibrium the "standard equilibrium" for the fixed contribution mechanism. There are other equilibria of this game, including the equilibrium where every type says "No." However, it is immediate that any equilibrium in weakly dominant strategies will be equivalent to the equilibrium described above for all $\mathbf{T}$ and all $t \in T$ where $\forall i$, $\hat{v}_i(t_i) \neq s_i$. That is, the equilibrium in weakly dominant strategies is unique up to the behavior of agents whose valuations exactly equal their share costs.

Given the equilibrium above, the direct mechanism version of the fixed contribution mechanism with shares s is of the form $F = (q^F(v), \tau^F(v))$, where

$$q^F(v) = I_{v \geq s} \tag{2.19}$$

where I_X is the indicator function that equals 1 if X is true and 0 otherwise, and

$$\tau_i^F(v) = s_i \cdot I_{v \geq s} \tag{2.20}$$

Lemma II.15 shows that this corresponds to a belief-invariant equilibrium of the mechanism. Therefore fixed contribution mechanisms are dominant strategy mechanisms.

Our interest in fixed contribution mechanisms derives from the following proposition. I restrict attention to deterministic dominant strategy mechanisms, i.e. where $q^M(t) \in \{0, 1\}$ although I conjecture that a similar result holds for stochastic dominant strategy mechanisms.

Proposition II.16. *For any belief-invariant equilibrium of a feasible mechanism,*

with the corresponding direct mechanism $M = (q^M(v), \tau^M(v))$ *with truth-telling equilibrium* a^**, there exists a fixed-contribution direct mechanism* $F = (q^F(v), \tau^F(v))$ *with truth-telling equilibrium* $\hat{a}^*$ *such that either:*

$$F \text{ and } \hat{a}^* \text{ improve on } M \text{ and } a^*$$

$$\text{or}$$

$$\forall v, \qquad q^M(v) = q^F(v) \text{ whenever } \sum_{i=1}^{N} v_i \neq c.$$

Proof. See appendix A. □

2.7 Additional Contribution Mechanism

This section describes a mechanism, called the "additional contribution mechanism," and discusses an equilibrium of this mechanism. Intuitively, this mechanism starts with fixed cost shares like the fixed contribution mechanism, but one agent can offer to pay a part of the fixed contribution of another agent (or agents) towards the production cost of the public good. Offering to pay part of another agent's payment will potentially make the other agent willing to support the public good's creation.

The additional contribution mechanism has the following structure. The mechanism designer sets initial shares $s^0 = (s_1^0, ..., s_N^0)$ with $\sum_{i=1}^{N} s_i^0 = c$. Without loss of generality, assume $s_1^0 < \overline{v}$.

The mechanism has two stages:

1. Each agent $i \leq N$ learns her type. Agent 1 selects a vector $b \in \mathbf{R}_+^N$ where for all i, $0 \leq b_i \leq s_i^0 - \underline{v}$, and $b_1 \equiv 0$. This choice corresponds to how much agent 1 offers to pay on behalf of each other agent (hence the component corresponding to agent 1 herself is zero.)

2. Each agent i is given a final cost share $\hat{s}_i$. For agent 1, that share is her initial share plus the sum of the components of b. For each agent $1 < i \leq N$, agent i's new share is her old share *minus* b_i, the corresponding component of b. Formally, the final cost shares $\hat{s}_i$ are determined by the following formula:

$$\hat{s}_1 = s_1^0 + \sum_{i=2}^{N} b_i$$

and for $i > 1$,

$$\hat{s}_i = s_i^0 - b_i$$

The agents play the fixed contribution mechanism with contribution shares $\hat{s}$.

It may seem counterintuitive that agent 1 would raise her own share by making an additional contribution to one or more other agents. However, if agent 1 believes that lowering another agent's required contribution increases the likelihood the public good is created by enough to make up for the increased private cost if the public good is created, then raising her own required contribution can be incentive compatible. I show an example where this occurs in section 2.8. Here I show certain properties that will hold for any equilibrium of the mechanism that is in non-weakly dominated strategies.

For the following analysis, fix a type space $\mathbf{T}$. A strategy for agent 1 is then a pair $(\beta_{\mathbf{T}}, m_{\mathbf{T},1})$ where $\beta_{\mathbf{T}} : T_1 \to \Delta(\mathbf{R}^N)$ maps every type $t_1 \in T_1$ into a distribution over b vectors, and where $m_{\mathbf{T},1} : T_1 \times \mathbf{R} \to \{\text{“Yes”, “No”}\}$ maps every type $t_1 \in T_1$ and observed final share $\hat{s}_1$ into an acceptance or rejection decision by agent 1. A strategy for an agent $1 < i \leq N$ is $m_{\mathbf{T},i} : T_i \times \mathbf{R} \to \{\text{“Yes”, “No”}\}$ which maps every type $t_i \in T_i$ and observed final share $\hat{s}_i$ into an acceptance or rejection decision by the agent. Technically, each agent's strategy in the final stage could depend on the $\hat{s}$'s or

β's, but because the fixed contribution mechanism is a dominant strategy mechanism, I can ignore the potential role of this information and focus on the dominant strategy equilibrium in the final stage.

Proposition II.17. *There exists an equilibrium a^* of the additional contribution mechanism such that*

$$\hat{v}(t) \geq s^0 \Rightarrow q^M(a^*_{\mathbf{T}}(t)) = 1 \tag{2.21}$$

and

$$\hat{v}(t) \geq \hat{s} \Rightarrow q^M(a^*_{\mathbf{T}}(t)) = 1 \tag{2.22}$$

for all $t \in T$ and all $\mathbf{T} \in \mathbf{\Omega}$.

Proof. I construct the equilibrium as follows: for all i, $m_{\mathbf{T},i}$ is specified by every agent plays the equilibrium of the fixed contribution mechanism described in section 6 in the second stage. This is enough to prove equation (2.22). To complete the description of the equilibrium we have to specify agent 1's choice $\beta_{\mathbf{T}}$ of a b vector.

If $\hat{v}_1(t_1) \leq s^0_1$: In this case $\beta_{\mathbf{T}}(t_1) = \vec{0}$. Agent 1 has no incentive to deviate because agent 1 will always get a utility of zero in the second stage regardless of the b agent 1 chooses, as $\hat{s}_1 \geq s^0_1 \geq \hat{v}_1(t_1)$. In the second stage, agent 1 will either be indifferent between "Yes" and "No" or will strictly prefer "No".

If $\hat{v}_1(t_1) > s^0_1$: In this case, $\beta_{\mathbf{T},-1}(t_1)$ maximizes the following expression:

$$\beta_{\mathbf{T},-1}(t_1) \in \underset{0 \leq b_{-1} \leq \vec{v} - s^0_{-1}}{\operatorname{argmax}} \hat{\pi}_{t_1}\left[s^0_{-1} - b_{-1} \leq v_{-1}(t_{-1})\right] \cdot \left(\hat{v}_1(t_1) - (s^0_1 + \sum_{i=2}^{N} b_i)\right). \tag{2.23}$$

Consider the function

$$F(b_{-1}) = \hat{\pi}_{t_1}\left[s^0_{-1} - b_{-1} \leq v_{-1}(t_{-1})\right] \cdot \left(\hat{v}_1(t_1) - (s^0_1 + \sum_{i=2}^{N} b_i)\right) \tag{2.24}$$

which is the expression within the argmax of equation (2.23) and is the payoff agent 1 expects for each possible b_{-1}. It is the product of $F_1(b_{-1})$ and $F_2(b_{-1})$, where

$$F_1(b_{-1}) = \hat{\pi}_{t_1}\left[s^0_{-1} - b_{-1} \leq v_{-1}(t_{-1})\right] = 1 - \hat{\pi}_{t_1}\left[v_{-1}(t_{-1}) < s^0_{-1} - b_{-1}\right] \tag{2.25}$$

which by inspection is everywhere non-negative, monotonically non-decreasing and right continuous in each dimension of b_{-1}, and $F_2(b_{-1})$ is the continuous function,

$$F_2(b_{-1}) = \hat{v}_1(t_1) - (s^0_1 + \sum_{i=2}^{N} b_i), \tag{2.26}$$

equal to the difference between agent 1's valuation and $\hat{s}_1$ as a function of b_{-1}, that is positive over some range.

I show $F(b_{-1})$ has a maximum by contradiction. Assume $F(b_{-1})$ has no maximum. Then there exists a sequence of $\{b^m_{-1}\}$ such that $F(b^m_{-1}) > 0$ is strictly increasing with m and there is no $\hat{b}_{-1}$ such that $F(\hat{b}_{-1}) \geq F(b^m_{-1})$ for all m. The domain of possible b_{-1} is compact, so the sequence b^m_{-1} has a convergent subsequence. Let the limit of this subsequence be b^*_{-1}. By F_1 monotonically non-decreasing and right-continuous in each dimension, $F_1(b^*_{-1}) \geq \lim F_1(b^m_{-1})$. And by F_2 continuous, $F_2(b^*_{-1}) = \lim F_2(b^m_{-1})$. $F(b^m_{-1}) > 0$ implies $F_2(b^m_{-1}) > 0$ for all m. Therefore, using that F_1 is everywhere non-negative and F_2 is positive for all b^m_{-1},

$$F(b^*_{-1}) = F_1(b^*_{-1}) \cdot F_2(b^*_{-1}) \geq \lim F_1(b^m_{-1}) \cdot \lim F_2(b^m_{-1}) = \lim F(b^m_{-1}) \tag{2.27}$$

and by assumption $F(b_{-1}^m)$ is strictly increasing in m, so for all m

$$F(b_{-1}^*) \geq F(b_{-1}^m). \tag{2.28}$$

However, that contradicts the assumption that $F(b_{-1})$ has no maximum. Therefore $F(b_{-1})$ must have a maximum and so the argmax in equation (2.23) is well defined.

If there is more than one argmax, then an argmax is chosen arbitrarily, constrained to

$$s_1^0 + \sum_{i=2}^{N} \beta_i < \hat{v}_1(t_1) \tag{2.29}$$

Choosing a $\beta(t_1)$ that satisfies equation (2.29) ensures that the first part of the Proposition holds. There must be a b that is an argmax of equation (2.23) and satisfies equation (2.29) for the following reason: either t_1's expected utility at the argmax b is greater than zero, implying that agent 1 expects the public good will be produced with some probability and that equation (2.29) holds (that is, t_1 pays less than $\hat{v}_1(t_1)$), or the expected utility at the argmax is 0, in which case $b = \vec{0}$ is an argmax. In the latter case let $\beta_1(t_1) = \vec{0}$. By assumption $\hat{v}_1(t_1) > s_1^0$ so equation (2.29) is satisfied.

Then equation (2.21) follows from equation (2.29) and the fact that for all $i > 1$ $\hat{s}_i \leq s_i^0$ by construction, as well as the equilibrium actions in the second stage of the game.

□

I will refer to an equilibrium of the additional contribution mechanism that fits the description of Proposition II.17 as a "standard equilibrium" of the additional contribution mechanism.

2.8 Comparison of Efficiency

In this section I compare the ex-post welfare of the standard equilibrium of the fixed contribution mechanism, as described in section 2.6, and a standard equilibrium of the additional contribution mechanism as described in section 2.7. The central result (Proposition II.22) is that the additional contribution mechanism performs at least as well as the fixed contribution mechanism on *any* type space, and performs better on some type spaces. In the terminology introduced in section 2.4, this means that any standard equilibrium of the additional contribution mechanism will improve on the standard equilibrium of the fixed contribution mechanism.

Our comparison will be between the fixed contribution mechanism with initial shares s^0 and the additional contribution mechanism with the same initial shares s^0, with the corresponding standard equilibria. The criterion will be the improvability relationship, as defined in section 2.4.

Proposition II.18. *For any type space* $\mathbf{T}$ *and type profile* t*, and given the standard equilibria, the additional contribution mechanism implements the social welfare maximizing outcome if the fixed contribution mechanism implements the social welfare maximizing outcome.*

Proof. There are two cases to consider: when welfare maximization requires that the public good not be created at t, and when welfare maximization requires the public good to be created.

When welfare maximization requires the public good not be created, the sum of valuations is less than c, so there is no possible $\hat{s}$ such that for all i, $\hat{s}_i \leq \hat{v}_i(t_i)$. Therefore in a standard equilibrium of the additional contribution mechanism, at least one agent must choose "No" and the public good will not be created. Likewise it must be that $s_i^0 > v_i$ for some i, guaranteeing that the public good is not created under

the standard equilibrium of the fixed contribution mechanism. So neither mechanism will produce the good.

When welfare maximization requires the public good be created, we need to show that if the fixed contribution mechanism leads to the public good being created, then so does the additional contribution mechanism. Recall that I defined a standard equilibrium of the additional contribution mechanism as having the following property (equation (2.21)):

$$\hat{v}(t) \geq s^0 \Rightarrow q^M(a^*_{\mathbf{T}}(t)) = 1$$

for all $t \in T$ and all $\mathbf{T} \in \mathbf{\Omega}$, which ensures that whenever the standard equilibrium of the fixed contribution mechanism produces the public good, so does the standard equilibrium of the additional contribution mechanism.

Therefore both when producing the public good is efficient and when it is not efficient, the standard equilibrium of the additional contribution mechanism maximizes welfare if the standard equilibrium of the fixed contribution mechanism maximizes welfare. □

Proposition II.19. *There exist* $\mathbf{T}$ *containing type profiles* t *where, for the standard equilibria of the additional contribution mechanism, the additional contribution mechanism implements welfare maximization but the fixed contribution mechanism, with its standard equilibrium, does not.*

Proof. To demonstrate the existence of type spaces that satisfy this condition, I use the following result that only depends on the primitives of $\mathbf{T}$. The logic here is that if agent 1 has a valuation above her initial share, knows the true valuations of the other agents with probability one, and those agents all have valuations at or below their initial shares (with at least one strictly below), then agent 1 can ensure the

public good is created by choosing appropriate additional contributions to direct to the other agents.

Lemma II.20. *For any type profile t such that*

- $\sum_{i=1}^{N} \hat{v}_i(t_i) > c$,
- $\hat{\pi}_{t_1}\left(\{\hat{t}_{-1} \mid \hat{v}_{-1}(\hat{t}_{-1}) = \hat{v}_{-1}(t_{-1})\}\right) = 1$,
- *for* $i > 1$, $\hat{v}_i(t_i) \leq s_i^0$, *and*
- *for some* $j \neq 1$, $\hat{v}_j(t_j) < s_j^0$

any standard equilibrium of the additional contribution mechanism implements the welfare maximizing outcome (creating the public good) at t.

Proof. Given agent 1's beliefs, agent 1 will only expect the public good to be created if and only if for all $i \neq 1$, $b_i \geq s_i^0 - \hat{v}_i(t_i)$. The minimum amount of additional contribution that will lead to the public good being created is then

$$\sum_{i \neq 1}^{N} s_i^0 - \hat{v}_i(t_i)$$

If agent 1 makes that additional contribution, she receives a payoff of

$$\hat{v}_1(t_1) - \hat{s}_i = \hat{v}_1(t_1) - s_1^0 - \sum_{i \neq 1}^{N} s_i^0 + \sum_{i \neq 1}^{N} \hat{v}_i(t_i)$$

$$= \hat{v}_1(t_1) - s_1^0 - \left(c - s_1^0\right) + \sum_{i=1}^{N} \hat{v}_i(t_i) - \hat{v}_1(t_1)$$

$$= \sum_{i=1}^{N} \hat{v}_i(t_i) - c > 0$$

Agent 1 gets a payoff of 0 if the public good is not created, so she is willing to pay the minimal additional contributions that ensure the public good gets created. Therefore, for all $i > 1$

$$b_i = s_i^0 - \hat{v}_i(t_i)$$

and for all i, $\hat{s}_i \leq \hat{v}_i(t_i)$ which implies the public good is created. □

Let $\mathbf{T}$ be any type space. Then the above lemma leads immediately to the following sufficient condition:

Lemma II.21. *If there exists a $t' \in T$ that satisfies the conditions of Lemma II.20, then at $t' \in T$ the additional contribution mechanism and its standard equilibria implement the welfare maximizing outcome (the public good is created), while the fixed contribution mechanism and its standard equilibrium do not.*

Proof. Lemma II.20 implies that the additional contribution mechanism implements the welfare maximizing outcome. The fixed contribution mechanism, however, will not lead to the public good being created, because there is some j such that $\hat{v}_j(t_j) < s_j^0$, so the fixed contribution mechanism does not implement the welfare maximizing outcome. □

Notably, the universal type space satisfies the conditions of lemma II.20, as do some finite type spaces. So if $\mathbf{\Omega}$ includes the set of all finite type spaces or the universal type space then lemma II.21 will hold for some $\mathbf{T} \in \mathbf{\Omega}$.

□

Propositions II.18 and II.19 together give us our result:

Proposition II.22. *If $\mathbf{\Omega}$ includes at least one $\mathbf{T}$ that satisifies the conditions of lemma II.20, then: for any shares s^0, the additional contribution mechanism with*

initial shares s^0 and its standard equilibrium improves on the fixed contribution mechanism with fixed shares $s = s^0$ and its standard equilibrium.

Proof. Proposition II.18 establishes that for every $\mathbf{T}$ and every $t \in T$, the additional contribution mechanism is at least as efficient. Proposition II.19 establishes that for some $\mathbf{T}$ and $t \in T$, the additional contribution mechanism is strictly more efficient. These two claims together establish that the additional contribution mechanism improves on the fixed contribution mechanism. □

Combined with Proposition II.16, which shows that any feasible dominant strategy mechanism is either improved on or equivalent (from an efficiency standpoint) to a fixed contribution mechanism, Proposition II.22 implies that every feasible dominant strategy mechanism is improved on by some additional contribution mechanism on any $\mathbf{\Omega}$ with a $\mathbf{T}$ that satisfies the conditions of lemma II.20. In particular, if $\mathbf{\Omega}$ includes the set of all finite type spaces or the universal type space, then every feasible dominant strategy mechanism is improved on by some additional contribution mechanism.

2.9 Conclusion

This paper has introduced, for the public goods problem, an approach to efficiency comparisons between mechanisms. One mechanism improves on another mechanism if it is at least as efficient for any realization of agents' type on any type space the mechanism designer considers possible, and strictly more efficient for some realization of agents' types on at least one of those type spaces. I demonstrated the soundness of the unimprovability concept on finite type spaces, and then used the improvability ranking to show that for any dominant strategy mechanism there exists an mechanism where agents' strategies depend upon their beliefs that improves on the dominant strategy mechanism, if the mechanism designer considers a sufficiently rich set of

type spaces. The result suggests than any welfare-maximizing mechanism designer who considers a sufficiently rich set of type spaces would prefer a mechanism other than a dominant strategy mechanism.

The improvability comparison raises the question of what mechanisms are unimprovable in the public good setting. Chapter 3 of this dissertation provides a first step in examining that question: a fairly simple mechanism, the all or nothing mechanism, is unimprovable on the universal type space among mechanisms with finite actions. I hope to characterize the (possibly large) set of undominated mechanisms on rich type spaces in the public good setting in future research. Another area for further work is extending the improvability concept and the associated analysis of mechanisms to other settings and to mechanism designers with objectives other than welfare maximization. Chapter 4 of this dissertation performs a similar analysis in the voting mechanism setting.

References

Bergemann, Dirk, and Morris, Stephen, 2005. "Robust Mechanism Design," Econometrica 73, 1771-1813.

Chung, Kim-Sau, and Ely, Jeffrey C., 2007. "Foundations of Dominant Strategy Mechanisms," Review of Economic Studies 74, 447-476.

Clarke, E, 1971. "Multipart Pricing of Public Goods," Public Choice 8, 19-33.

d'Aspremont, Claude, and Gerard-Varet, Louis-Andre, 1979. "Incentives and Incomplete Information," Journal of Public Economics 11, 24-45.

Groves, T., 1973. "Incentives in Teams," Econometrica 41, 617-631.

Güth, Werner, and Hellwig, Martin, 1986. "The Private Supply of a Public Good," Journal of Economics, Supplement 5, 121-159.

Ledyard, John O., and Palfrey, Thomas R., 2007. "A General Characterization of Interim Efficient Mechanisms for Independent Linear Environments," Journal of Economic Theory 133, 441-466.

Mailath, George J., and Postlewaite, Andrew, 1990. "Asymmetric Information Bargaining Problems with Many agents," Review of Economic Studies 57, 351-367.

Mertens, Jean-Francois, and Zamir, Shmuel, 1985. "Formulation of Bayesian Analysis for Games with Incomplete Information," International Journal of Game Theory 14, 1-29.

Varian, H. R.,1994a. "Sequential Contributions to Public Goods," Journal of Public Economics 53, 165 - 186.

Varian, H.R., 1994b. "A Solution to the Problem of Externalities When Agents are Well Informed," American Economic Review, 84, 1278-1293.

Vickrey, W.,1961. "Counterspeculation, Auctions and Competitive Sealed Tenders," Journal of Finance 16, 8-37.

Wilson, Robert, 1987. "Game-Theoretic Analyses of Trading Processes," *Advances in Economic Theory: Fifth World Congress*, ed. by T. Bewley. Cambridge, U.K.: Cambridge University Press, Chapter 2, 33-70.

Yamashita, Takuro, 2011. "Robust Welfare Guarantees in Bilateral Trading Mechanisms," discussion paper, Stanford University.

2.10 Appendix 2.A

Dominant Strategy Mechanisms Are (Weakly) Improved on by Fixed Contribution Mechanisms

The following lemma is a standard result. I present its proof for completeness.

Lemma II.23. *A mechanism that satisfies ex-post budget balance (or has a non-negative budget balance everywhere) is dominant strategy incentive compatible and ex-post individual rational if and only if, for all i there exists a function $\tau_i(v_{-i})$ such that, for all v_{-i},*

$$q(\hat{v}_i, v_{-i}) = \begin{cases} 1 & \hat{v}_i > \tau_i(v_{-i}) \\ 0 & \hat{v}_i < \tau_i(v_{-i}) \end{cases} \tag{2.30}$$

and

$$\tau_i(\hat{v}_i, v_{-i}) = \begin{cases} \tau_i(v_{-i}) & \hat{v}_i > \tau_i(v_{-i}) \\ 0 & \hat{v}_i < \tau_i(v_{-i}) \end{cases} \tag{2.31}$$

Proof. Consider any v_{-i}.

If $\forall \hat{v}_i$, $q(\hat{v}_i, v_{-i}) = 0$, then let $\tau_i(v_{-i}) = \overline{v}$.

For any $\hat{v}_i$ and $\hat{v}_i'$ such that $q(\hat{v}_i, v_{-i}) = q(\hat{v}_i', v_{-i}) = 1$, then dominant strategy incentive compatible implies that $\tau_i(\hat{v}_i, v_{-i}) = \tau_i(\hat{v}_i', v_{-i})$. Therefore let $\tau_i(v_{-i}) \equiv \tau_i(\hat{v}_i, v_{-i})$ for any $\hat{v}_i$ such that $q(\hat{v}_i, v_{-i}) = 1$.

If there exist $\hat{v}_i$ such that $q(\hat{v}_i, v_{-i}) = 0$, then by ex-post individual rationality and that the budget cannot be negative, $\tau_i(\hat{v}_i, v_{-i}) = 0$.

Then by dominant strategy incentive compatibility,

$$q(\hat{v}_i, v_{-i}) = 1 \quad \Rightarrow \quad \hat{v}_i \geq \tau_i(v_{-i}) \quad \text{and} \quad \tau_i(\hat{v}_i, v_{-i}) = \tau_i(v_{-i}) \tag{2.32}$$

and

$$q(\hat{v}_i, v_{-i}) = 0 \quad \Rightarrow \quad \hat{v}_i \leq \tau_i(v_{-i}) \quad \text{and} \quad \tau_i(\hat{v}_i, v_{-i}) = 0 \tag{2.33}$$

which imply equation (2.30) and equation (2.31).

□

Proposition II.16. *For any belief-invariant equilibrium of a feasible mechanism, with the corresponding direct mechanism $M = (q^M(v), \tau^M(v))$ with truth-telling equilibrium a^*, there exists a fixed-contribution direct mechanism $F = (q^F(v), \tau^F(v))$ with truth-telling equilibrium $\hat{a}^*$ such that either:*

$$F \text{ and } \hat{a}^* \text{ improve on } M \text{ and } a^*$$

$$\text{or}$$

$$\forall v, \qquad q^M(v) = q^F(v) \text{ whenever } \sum_{i=1}^{N} v_i \neq c.$$

Proof. In constructing F, there are two possibilities to consider regarding $q^M(\overline{v})$:

If $q^M(\overline{v}) = 0$ then the result is trivial. Set $s_i = \frac{c}{N}$ for all i, and then it is immediate that F and $\hat{a}^*$ improve on M and a^*.

Therefore the rest of the proof will deal with the case $q^M(\overline{v}) = 1$.

Construct $F = (q^F(v), \tau^F(v))$ by, for all i, setting $s_i = \tau_i^M(\overline{v})$.

Then $q^F(v) = I_{v \geq s}$ and $\tau_i^F(v) = q^F(v) \cdot s_i$.

F is feasible because the fact that M is budget balanced implies

$$\sum_{i=1}^{N} s_i = \sum_{i=1}^{N} \tau_i(\overline{v}) = 1.$$

For any v, we can compare the relative efficiency.

Case 1: $\sum_{i=1}^{N} v_i < c$. In this case creation of the public good is inefficient. Furthermore, $\sum_{i=1}^{N} v_i < c = \sum_{i=1}^{N} s_i$ so there exists an i such that $v_i < s_i$ which implies $q^F(v) = 0 \leq q^M(v)$.

Case 2: $\sum_{i=1}^{N} v_i > c$. There are two possible values for $q^M(v)$:
If $q^M(v) = 1$, note that lemma II.23 implies that $q(v_i, v_{-i})$ is non-decreasing in v_{-i}. That implies that $q(v_j, \overline{v}_{-j}) = 1$ for all j. Lemma II.23 then implies that $v_j \geq \tau_j^M(\overline{v}) = s_j$ for all j. Therefore $q^F(v) = 1 = q^M(v)$.

If $q^M(v) = 0$, then $q^F(v) \geq 0 = q^M(v)$. Therefore $q^F(v) \geq q^M(v)$.

Together these results show that in Case 2, $q^F(v) \geq q^M(v)$.

Combining Case 1 and Case 2, we get the statement:

$$\begin{aligned} q^F(v) \geq q^M(v) \quad & \text{when } \sum_{i=1}^{N} v_i > c \\ & \text{and} \\ q^F(v) \leq q^M(v) \quad & \text{when } \sum_{i=1}^{N} v_i < c \end{aligned} \tag{2.34}$$

If there exists a v such that the applicable inequality in (2.34) is strict, then F and $\hat{a}^*$ improve on M and a^*; otherwise, (2.34) implies that for all v,

$$\sum_{i=1}^{N} v_i \neq c \Rightarrow q^F(v) = q^M(v)$$

which completes the result to be proved.

□

2.11 Appendix 2.B

Non-Negative Budget Balance

In the main body of this paper I have assumed that feasible mechanisms have a strict budget balance of zero. A weaker assumption would be to allow mechanisms to have a non-negative budget balance. Here I discuss the implications of a non-negative budget balance for the definition of improvability, and for the improvability of dominant strategy mechanisms by fixed contribution mechanisms. I show that allowing non-negative budget balance does not alter the result of the paper regarding the improvability of dominant strategy mechanisms, as fixed contribution mechanisms still (weakly) improve on all other non-negative budget balance dominant strategy mechanisms.

Improvability. Allowing non-negative budget balance means that the definition of improvability has to be modified. Now it must account for the fact that the sum of transfers affects the sum of the individual agents' utilities as well as whether the public good is produced. How to define improvability will now depend on what happens to the non-balanced portion of transfers.

If we assume the budget is balanced by an agent outside of the N agents we've modeled as participating in the mechanism, *and the mechanism designer cares equally about this external agent's welfare as the welfare of the participants*, then the definition of improvability used above still applies; any transfer paid by an agent is received by another agent, except for transfers used to pay for the public good, so a mechanism maximizes welfare by creating the public good if its aggregate value exceeds its cost and not creating the public good if its aggregate value is less than the cost.

Alternatively we might imagine that a mechanism constrained to have a non-negative budget might generate a surplus that is either physically destroyed or transferred to agents whose welfare does not enter into the mechanism designer's consideration. I call such a mechanism a non-negative budget mechanism.

Definition II.24. A mechanism $M = (A, \hat{y})$ is a **non-negative budget mechanism** if for all $a \in A$,

$$\sum_{i=1}^{N} \tau_i(a) \geq c \cdot q(a) \tag{2.35}$$

Then the definition of improvability can be generalized to become the following definition of improvability-including-transfers:

Definition II.25. A non-negative budget mechanism $M = (q^M, \tau^M)$ and equilibrium a^* **improve-including-transfers on** another non-negative budget mechanism $\hat{M} = (q^{\hat{M}}, \tau^{\hat{M}})$ and equilibrium $\hat{a}^*$ if for all type spaces $\mathbf{T} \in \mathbf{\Omega}$ and all $t \in T$,

$$q^M(a^*_{\mathbf{T}}(t)) \sum_{i=1}^{N} \hat{v}_i(t_i) - \sum_{i=1}^{N} \tau_i^M(a^*_{\mathbf{T}}(t)) \geq q^{\hat{M}}(\hat{a}^*_{\mathbf{T}}(t)) \sum_{i=1}^{N} \hat{v}_i(t_i) - \sum_{i=1}^{N} \tau_i^{\hat{M}}(\hat{a}^*_{\mathbf{T}}(t)) \tag{2.36}$$

and there is some $\mathbf{T} \in \mathbf{\Omega}$ and $t \in T$ such that the inequality is strict.

By inspection this definition corresponds to (weakly) increasing the sum of the realized utilities of the agents at each possible type profile. If both mechanisms are ex-post budget balanced in the strict sense of equation (3), then improvability-with-transfers is equivalent to improvabilty.

Dominant strategy mechanisms. In this part of the appendix I analyze dominant strategy mechanisms that satisfy non-negative budget balance and individual rationality but not necessarily ex-post budget balance.

I show that proposition 1 generalizes to non-negative budget, individually rational dominant strategy mechanisms with non-negative budget balance (Proposition II.26). I further show that proposition 1 generalizes to non-negative budget, individually rational dominant strategy mechanisms when the criteria is changed from improvability to improvability-with-transfers (Proposition II.28).

Proposition II.26. *For any belief-invariant equilibrium of a non-negative budget, individually rational mechanism, with the corresponding direct mechanism* $M = (q^M(v), \tau^M(v))$ *with truth-telling equilibrium* a^**, there exists a fixed-contribution direct mechanism* $F = (q^F(v), \tau^F(v))$ *with truth-telling equilibrium* $\hat{a}^*$ *such that either:*

$$F \text{ and } \hat{a}^* \text{ improve on } M \text{ and } a^*$$

$$\text{or}$$

$$\forall v, \qquad q^M(v) = q^F(v) \text{ whenever } \sum_{i=1}^{N} v_i \neq c.$$

Proof. This section draws upon the proof for Proposition 1, in Appendix A. Note that for the proof of Proposition 1 the only detail of a particular dominant strategy mechanism M that matters is $\tau(\overline{v})$. Lemma II.23 only requires that the budget cannot be negative.

Fix a non-negative budget, individually rational dominant strategy mechanism $M = (q^M(v), \tau^M(v))$. If $\sum_{i=1}^{N} \tau_i(\overline{v}) = c$, then the proof of Proposition 1 applies without any modification.

If $\sum_{i=1}^{N} \tau_i^M(\overline{v}) > c$, then proposition 1 only needs to be modified in a very small way: in the construction of F, let

$$s_1 = \tau_i^M(\overline{v}) - \left(\sum_{i=1}^{N} \tau_i^M(\overline{v}) - c \right)$$

That is, in the fixed contribution mechanism that is constructed, I lower agent 1's fixed share by just enough to balance the budget (at $\overline{v}$). Then the proof of proposition 1 applies.

□

Note that in the case that $\sum_{i=1}^{N} \tau_i^M(\overline{v}) > c$ the fixed contribution mechanism must improve on (not just be equivalent to) the original mechanism, because the fixed contribution mechanism creates the public good for any v of the form $(v_1, \overline{v}_{-1})$, where

$$v_1 \in \left[\tau_i^M(\overline{v}) - \left(\sum_{i=1}^{N} \tau_i^M(\overline{v}) - c \right),\ \tau_i^M(\overline{v}) \right)$$

but the original mechanism does not.

Therefore the result of Proposition 1 extends to the entire class of non-negative budget, individually rational dominant strategy mechanisms (of which feasible dominant strategy mechanisms are a subset).

As the following lemma demonstrates, it is straighforward to show that a strictly budget balanced mechanism that improves on a non-negative budget mechanism also improves-including-transfers on the non-negative budget mechanism.

Lemma II.27. *If an ex-post budget balanced mechanism $M = (q^M, \tau^M)$ and equilibrium a^* improve on a non-negative budget mechanism $\hat{M} = (q^{\hat{M}}, \tau^{\hat{M}})$ and equilibrium*

$\hat{a}^*$, *then*

$$M \text{ and } a^* \text{ improve-including-transfers on } \hat{M} \text{ and } \hat{a}^*.$$

Proof. For any type space $\mathbf{T} \in \mathbf{\Omega}$ and $t \in T$,

M ex-post budget balanced and $\hat{M}$ non-negative budget implies:

$$\sum_{i=1}^{N} \tau_i^M(a^*_{\mathbf{T}}(t)) = c \cdot q^M(a^*_{\mathbf{T}}(t)) \quad \text{and} \quad \sum_{i=1}^{N} \tau_i^{\hat{M}}(\hat{a}^*_{\mathbf{T}}(t)) \geq c \cdot q^{\hat{M}}(\hat{a}^*_{\mathbf{T}}(t)) \tag{2.37}$$

Improvability implies:

$$q^M(a^*_{\mathbf{T}}(t)) \left(\sum_{i=1}^{N} \hat{v}_i(t_i) - c \right) \geq q^{\hat{M}}(\hat{a}^*_{\mathbf{T}}(t)) \left(\sum_{i=1}^{N} \hat{v}_i(t_i) - c \right)$$

Substituting from (2.37) yields

$$q^M(a^*_{\mathbf{T}}(t)) \sum_{i=1}^{N} \hat{v}_i(t_i) - \sum_{i=1}^{N} \tau_i^M(a^*_{\mathbf{T}}(t)) \geq q^{\hat{M}}(\hat{a}^*_{\mathbf{T}}(t)) \sum_{i=1}^{N} \hat{v}_i(t_i) - \sum_{i=1}^{N} \tau_i^{\hat{M}}(\hat{a}^*_{\mathbf{T}}(t))$$

which is identical to equation (2.36). Therefore M and a^* improve-including-transfers on $\hat{M}$ and $\hat{a}^*$. □

Proposition II.26 and lemma II.27 imply a generalization of Proposition II.16.

Proposition II.28. *For any belief-invariant equilibrium of a non-negative budget, individually rational mechanism, with the corresponding direct mechanism* $M = (q^M(v), \tau^M(v))$ *with truth-telling equilibrium* a^**, there exists a fixed-contribution direct mechanism* $F = (q^F(v), \tau^F(v))$ *with truth-telling equilibrium* $\hat{a}^*$ *such that either:*

F *and* $\hat{a}^*$ *improve-with-transfers on* M *and* a^*

or

$$\forall v, \qquad q^M(v) = q^F(v) \text{ whenever } \sum_{i=1}^{N} v_i \neq c.$$

CHAPTER III

An Unimprovability Result for the Public Good Problem

3.1 Introduction

Chapter 2 analyzed the problem of designing a mechanism for the public good problem when the mechanism designer does not make assumptions about agents' types, but instead considers a wide range of type spaces as plausible descriptions of agents' preferences and beliefs. I looked specifically at the approach of finding a weakly undominated (or "unimprovable") mechanism when considering performance across a range of type spaces. I showed that mechanisms exist that are unimprovable on finite type spaces. However, these mechanisms may be too complicated in practice to be of much use to a mechanism designer. Consequently it may be useful to restrict attention to a smaller set of mechanisms that a mechanism designer could realistically implement.

In this paper I attempt such an analysis by focusing on mechanisms with finite actions. I also take the relevant concept of improvability to be that a mechanism is improvable if there exists a mechanism $\hat{M}$ with finite actions that improves on M. To model the uncertainty of the mechanism designer regarding the preferences and beliefs of agents, I assume the mechanism designer considers outcomes on the

universal type space.

In this paper I demonstrate a finite action mechanism, the All or Nothing mechanism, that is unimprovable among the set of mechanisms with finite actions. I first describe the mechanism and then provide the result regarding unimprovability. A weakness of the result is that the method of proof is not particularly generalizable to other mechanisms. However, the result does show both that a relatively simple mechanism can be unimprovable, and that such a result can be proven using relatively unsophisticated (if involved) methods.

The paper is organized as follows. Section 3.2 briefly summarizes the public good environment (described in more detail in the companion paper) and highlights the differences between the setting in this paper and the companion paper. Section 3.3 presents the All or Nothing mechanism. Section 3.4 establishes some useful lemmas for proving the main result. Section 3.5 contains the proof that the All or Nothing mechanism is unimprovable. Section 3.6 concludes.

3.2 The Mechanism Design Problem

In general I follow the assumptions and notations of chapter 2; I refer the reader to that document for definitions and assumptions. However, I make some changes to the environment for this paper. I assume that $\overline{v} > c$, that is the highest possible valuation is greater than the cost of the public good. For technical reasons, I also assume that mechanisms have a finite number of actions. Furthermore I focus on the case that there are exactly two agents, i.e. $N = 2$.

I assume that the mechanism designer's uncertainty about agents' preferences and beliefs can be captured by the mechanism designer considering outcomes over the universal type space. In the notation of chapter 2, I assume that $\mathbf{\Omega} = \{\mathbf{S}\}$. This simplifies description of equilibria by not requiring the equilibrium to be defined over

more than one type space.

I restrict attention to pure-strategy equilibria in non-weakly dominated strategies. I make one further non-standard restriction on the equilibria under consideration: in equilibrium, every action is *somewhere strictly preferred*. By this I mean for each agent, and each action of that agent played in equilibrium, there exists a type of that agent in the universal type space for whom that action is a strict best response.

Definition III.1. An equilibrium a^* satisfies *every action played in equilibrium is somewhere strictly preferred* on $\{\mathbf{S}\}$ if for all i and all $t_i \in T_i$, where T_i is the typespace for agent i in the universal typespace, there exists $\hat{t}_i$ such that

$$U_i(\hat{t}_i, a_i^*(\hat{t}_i)) > U_i(\hat{t}_i, a_i') \tag{3.1}$$

for all $a_i' \neq a_i^*(t_i)$.

3.3 The All or Nothing Mechanism

I assume in the following discussion that there are two agents, 1 and 2. I normalize the cost of the public good to 1, and assume the lowest possible valuation $\underline{v} = 0$. I furthermore assume $\overline{v} > 1$. The All or Nothing Mechanism is a simplified "take it or leave it" offer by agent 1, where agent 1 must either offer to pay the whole cost or force agent 2 to pay the entire cost if the public good is built. Alternatively, this mechanism is equivalent to a simplified additional contribution mechanism with initial shares $s_1^0 = 0$ and $s_2^0 = 1$, where the first agent can raise her own cost share, but only has the option of raising that share to the entire cost. (See chapter 2 for discussion of the additional contribution mechanism.)

Formally, The All or Nothing Mechanism is $\tilde{M} = \left\{(\tilde{A}_1, \tilde{A}_2), \overline{y}\right\}$ where

- $\tilde{A}_1 = \{$"PAY", "DON'T PAY"$\}$

- $\tilde{A}_2 = \{\text{“pay”}, \text{“don't pay”}\}$
- $\tilde{y}(a_1, a_2) = (\tilde{q}(a_1, a_2), \tilde{\tau}(a_1, a_2))$ is defined by the values in table 1.

$\tilde{q}(a_1, a_2) =$		
Actions	pay	don't pay
PAY	1	1
DON'T PAY	1	0

$\tilde{\tau}(a_1, a_2) =$		
Actions	pay	don't pay
PAY	(1,0)	(1,0)
DON'T PAY	(0,1)	(0,0)

Table 3.1: Outcomes under the All or Nothing Mechanism

The following strategies are an equilibrium of the All or Nothing Mechanism.

$$\tilde{a}_1^*(t_1)) = \begin{cases} \text{“PAY”} & \text{if } \hat{v}_1(t_1) > 1, \\ & \hat{\pi}_{t_1}\left[\hat{v}_2(t_2) < 1\right] \geq \frac{1}{\hat{v}_1(t_1)} \\ \text{“DON'T PAY”} & \text{otherwise} \end{cases}$$

$$\tilde{a}_2^*(t_2)) = \begin{cases} \text{“pay”} & \text{if } \hat{v}_2(t_2) \geq 1 \\ \text{“don't pay”} & \text{if } \hat{v}_2(t_2) < 1 \end{cases}$$

Lemma III.2. *The strategies $\tilde{a}_1^*(t_1)$ and $\tilde{a}_2^*(t_2)$ are an equilibrium of the All or Nothing mechanism on the universal type space.*

Proof. For agent 2, inspection shows that “pay” is a weakly dominant strategy when $\hat{v}_2(t_2) > 1$ and “don't pay” is a weakly dominant strategy when $\hat{v}_2(t_2) < 1$. When $\hat{v}_2(t_2) = 1$ then agent 2 is indifferent between the two actions regardless of his beliefs.

Therefore agent 2 has no incentive to deviate from $\tilde{a}_2^*(t_2))$ regardless of agent 1's strategy.

Given agent 2's strategy is $\tilde{a}_2^*(t_2)$, agent 1's payoff for each action correspond to:

$$U(t_1, \text{“PAY”}) = \hat{v}_1(t_1) - 1$$

and

$$U(t_1, \text{“DON'T PAY”}) = \hat{\pi}_{t_1}[\hat{v}_2(t_2) \geq 1] \cdot \hat{v}_1(t_1)$$

Therefore, the action “DON'T PAY” has a higher payoff for agent 1 when

$$U(t_1, \text{“DON'T PAY”}) \geq U(t_1, \text{“PAY”})$$

which is equivalent to

$$\hat{\pi}_{t_1}[\hat{v}_2(t_2) \geq 1] \cdot \hat{v}_1(t_1) \geq \hat{v}_1(t_1) - 1$$

which is equivalent to

$$\hat{\pi}_{t_1}[\hat{v}_2(t_2) < 1] \geq \frac{1}{\hat{v}_1(t_1)}.$$

Therefore $\tilde{a}_1^*(t_1)$ specifies that agent 1 play the action “DON'T PAY” exactly when is it incentive compatible to do so.

□

3.4 The Focus Lemma

This section introduces a result that will be used repeatedly in the proof of Proposition 1. It will be useful therefore to state and prove the result before proceeding to

the main proof. Let $i \in \{1,2\}$ and $j \neq i$.

Definition III.3. Given a set of distributions $\{\tilde{\pi}_i^n\}_{i=1,...,N}$ such that $\tilde{\pi}_i^n \in \Delta(T_j)$ for all n, the set $B_i(\{\tilde{\pi}_i^n\}, v_i) \in A_i$ is defined by

- $B_i(\{\tilde{\pi}_i^1\}, v_i) \subset A_i$ is the set of equilibrium best responses to $\tilde{\pi}_i^1$.
- For $m \in \{2,3,...,N\}$, $B_i(\{\tilde{\pi}_i^n\}_{n=1,...,m}, v_i) \subset B_i(\{\tilde{\pi}_i^n\}_{n=1,...,m-1}, v_i)$ is the set of actions that are best responses to $\tilde{\pi}_i^m$ conditional on being in $B_i(\{\tilde{\pi}_i^n\}_{n=1,...,m-1}, v_i)$.

Finiteness of A_i ensures that $B_i(\{\tilde{\pi}_i^n\}, v_i) \in A_i$ is non-empty. I now prove the following useful result, which I refer to as the "focus lemma."

Lemma III.4 (focus lemma)**.** *If for some* $\{\tilde{\pi}_i^n\}_{n=1,...,N}$ *and* $\vec{\epsilon} = (\epsilon_1, ..., \epsilon_N)$*,*

$$\hat{\pi}_i(t_i)[E] = \sum_{n=1}^{N} \epsilon_n \tilde{\pi}_i^n[E] \tag{3.2}$$

and

$$\overline{\epsilon} \equiv \max_{1 \leq n < N} \frac{\epsilon_{n+1}}{\epsilon_n} < 1 \tag{3.3}$$

Then $\overline{\epsilon}$ *sufficiently small implies* $a_i^*(t_i) \in B_i(\{\tilde{\pi}_i^n\}_{n=1,...,N}, \hat{v}_i(t_i))$.

Proof. By induction on n.

- $n = 1$. Because A_i is finite, there is some $\kappa_1 > 0$ such that for any action $\hat{a}_i \notin B_i(\{\tilde{\pi}_i^1\}, \hat{v}_i(t_i))$, conditional on the distribution $\tilde{\pi}_i^1$ the difference between the expected payoff of any action $a_i \in B_i(\{\tilde{\pi}_i^1\}, \hat{v}_i(t_i))$ and the expected payoff of $\hat{a}_i$ is at least κ_1. Because A is finite, the set of possible realized utilities for agent i (condition on $\hat{v}_i(t_i,)$) is also finite; let λ designate the difference between the maximal and the minimal possible realized utilities. Then the difference in expected utility for playing $\hat{a}_i$ versus a_i is bounded by

$$U_i(t_i, \hat{a}_i) - U_i(t_i, a_i) \leq \epsilon_1(-\kappa_1) + (1 - \epsilon_1) \cdot \lambda \tag{3.4}$$

Note that neither κ_1 nor λ depend on the choice of $\vec{\epsilon}$. Equation (3.3) implies that

$$1 - \epsilon_1 = \sum_{n=2}^{N} \epsilon_n \leq (N-1)\overline{\epsilon} \cdot \epsilon_1. \tag{3.5}$$

subsituting into equation (3.4) yields

$$U_i(t_i, \hat{a}_i) - U_i(t_i, a_i) \leq \epsilon_1(-\kappa_1) + ((N-1)\overline{\epsilon} \cdot \epsilon_1) \cdot \lambda \tag{3.6}$$

which implies

$$-\kappa_1 + (N-1)\overline{\epsilon} \cdot \lambda < 0 \quad \Rightarrow \quad U_i(t_i, \hat{a}_i) - U_i(t_i, a_i) < 0 \tag{3.7}$$

and consequently, if

$$\overline{\epsilon} < \frac{\kappa_1}{(N-1)\lambda} \tag{3.8}$$

then $U_i(t_i, \hat{a}_i) < U_i(t_i, a_i)$ for all $\hat{a}_i \notin B_i(\{\tilde{\pi}_i^1\}, \hat{v}_i(t_i))$ and therefore $a_i^*(t_i) \in B_i(\{\tilde{\pi}_i^1\}, \hat{v}_i(t_i))$.

- $n > 1$. By the inductive hypothesis, $a_i^*(t_i) \in B_i(\{\tilde{\pi}_i^m\}_{m=1,\ldots,n-1}, \hat{v}_i(t_i))$. I now look at the payoff of playing any action $\hat{a}_i \in B_i(\{\tilde{\pi}_i^m\}_{m=1,\ldots,n-1}, \hat{v}_i(t_i)) \setminus B_i(\{\tilde{\pi}_i^m\}_{m=1,\ldots,n}, \hat{v}_i(t_i))$ and an action $a_i \in B_i(\{\tilde{\pi}_i^m\}_{m=1,\ldots,n}, \hat{v}_i(t_i))$. As in the $n = 1$ case, let $\kappa_n > 0$ be the minimal difference between the expected payoff of any such a_i and the expected payoff of any such action $\hat{a}_i$, conditional on the distribution $\tilde{\pi}_i^n$. (If no such $\hat{a}_i$ exist then the inductive step is immediately satisfied). Note that κ_n only depends on $\{\tilde{\pi}_i^m\}_{m=1,\ldots,n}$. The difference in payoffs must be bounded by

$$U_i(t_i, \hat{a}_i) - U_i(t_i, a_i) \leq \epsilon_n(-\kappa_n) + \left(\sum_{m=n+1}^{N} \epsilon_m \right) \cdot \lambda \tag{3.9}$$

Note that equation (3.3) implies that $\sum_{m=n+1}^{N} \epsilon_m < (N-n)\overline{\epsilon} \cdot \epsilon_n$. Then a simple calculation similar to the $n = 1$ case shows that the right side of equation (3.9) is negative if

$$\overline{\epsilon} < \frac{\kappa_n}{(N-n)\lambda} \tag{3.10}$$

Therefore if equation (3.10) is satisfied for all $n \leq N$, then the result holds. This completes the proof. □

3.5 The All or Nothing Mechanism is Unimprovable

Now I can state the proposition.

Proposition III.5. *Given* $\mathbf{\Omega} = \{\mathbf{S}\}$*, the universal type space, the All or Nothing mechanism, with the equilibrium described above, is not improvable by any mechanism with finite actions and a pure-strategy equilibrium in non-weakly dominated strategies where every action played in equilibrium is somewhere strictly preferred.*

Proof. First I show that another proposition, proposition III.7, is equivalent to proposition III.5. I then prove proposition III.7 to complete the proof.

The proof takes into account the following observations regarding what it would mean for a mechanism and its equilibrium to improve on the All or Nothing mechanism. Another way to state the claim is that if there is a mechanism $M = (A, \hat{y})$ with equilibrium a^* that is as efficient for any realization of agents' types t as the All or Nothing mechanism $\tilde{M}$ with equilibrium $\tilde{a}^*$, then for all realizations of agents' types t, M and $\tilde{M}$ must be equally efficient. The proof focuses on this phrasing of the claim.

M as efficient as $\tilde{M}$: The All or Nothing mechanism is efficient whenever $\hat{v}_2(t_2) \geq 1$ or $\tilde{a}_1^*(t_1) =$"PAY". In the first case, the public good is produced because $\tilde{a}_2^*(t_2) =$"pay",

and because $\hat{v}_2(t_2) \geq 1$ the sum of the agents' valuations is at least one. In the second case, the public good is produced and $\tilde{a}_1^*(t_1) =$"PAY" implies that $\hat{v}_1(t_1) \geq 1$ so the sum of the agents' valuations are at least one. For the first case, M and a^* as efficient as $\tilde{M}$ and $\tilde{a}^*$ implies that for all

$$a_2 \in \{a_2 \in A_2 | \exists t_2 \in T_2,\ \hat{v}_2(t_2) \geq 1,\ a_2^*(t_2) = a_2\},$$

$$q(a_1, a_2) = 1 \quad \forall a_1 \in A_1 \tag{3.11}$$

that is, for any action that is played in equilibrium by some t_2 with a valuation at least 1, it must be that the public good is produced when that action is played, regardless of the action of agent 1. That implies that $\forall a_1 \in A_1, \forall t_1 \in T_1$,

$$Q_1(t_1, a_1) \geq \hat{\pi}_1(t_1)[\hat{v}_2(t_2) \geq 1]. \tag{3.12}$$

Similarly for the second case, for any

$$a_1 \in \{a_1 \in A_1 | \exists t_1 \in T_1,\ \tilde{a}_1^*(t_1) = \text{"PAY"},\ a_1^*(t_1) = a_1\},$$

$$q(a_1, a_2) = 1 \quad \forall a_2 \in A_2$$

that is, for any action played in equilibrium by a type of agent 1 who would play "PAY" under the equilibrium of the All or Nothing mechanism, it must be that when that action is played the public good is produced regardless of the action of agent 2.

The All or Nothing mechanism is also efficient when $\hat{v}_1(t_1) + \hat{v}_2(t_2) < 1$, that is when producing the public good is inefficient. In this situation $\tilde{a}^*$ dictates that agent 1 plays "DON'T PAY" and agent 2 plays "don't pay", as both agents' valuations are less than one.

How M could be more efficient than $\tilde{M}$: The All or Nothing mechanism is not efficient in the remaining situations, where $\hat{v}_2(t_2) < 1$ and $\tilde{a}_1^*(t_1) =$"DON'T PAY", but $\hat{v}_1(t_1) + \hat{v}_2(t_2) > 1$. If M and a^* improve on $\tilde{M}$ and $\tilde{a}^*$, it must be for some t that fits this description. Therefore to prove the proposition, I need to show the following:

Proposition III.6. *If $M = (A, \hat{y})$ and a^* are as efficient as $\tilde{M}$ and $\tilde{a}^*$ for all t, then for all t such that*

- $\hat{v}_2(t_2) < 1$, *and*
- $\tilde{a}_1^*(t_1) =$ *"DON'T PAY",*

it must be that in equilibrium a^* of M,

$$q(a_1^*(t_1), a_2^*(t_2)) = 0. \tag{3.13}$$

Given the observations above about when the All or Nothing mechanism is efficient, this proposition says that in those situations where the All or Nothing mechanism is not efficient, neither is M.

Expectations and ex-post improvement. Proposition III.6 looks at ex-post outcomes, but it will be useful to rephrase the required result in terms of agents' expectations. Define

$$Q_1(t_1) \equiv E_{\hat{\pi}_1(t_1)} \left[q(a_1^*(t_1), a_2^*(t_2)) \right]$$

and

$$\tilde{Q}_1(t_1) \equiv E_{\hat{\pi}_1(t_1)} \left[\tilde{q}(\tilde{a}_1^*(t_1), \tilde{a}_2^*(t_2)) \right].$$

Clearly if proposition III.6 holds then $Q_1(t_1) = \tilde{Q}_1(t_1)$ for all t_1. What we will take advantage of is the opposite fact, that if $Q_1(t_1) = \tilde{Q}_1(t_1)$ for all t_1, proposition III.6 holds. It is more complicated to show that $Q_1(t_1) = \tilde{Q}_1(t_1)$ for all t_1 implies proposition III.6. To do so, we observe that by the assumption that every action played in equilibrium is somewhere strictly preferred, for all $a_1 \in A_1$ there exists a type t_1 such that $a^*(t_1) = a_1$ and t_1 strictly prefers a_1 to any other action, in equilibrium. Then there exists some $\hat{\pi}_1' \in \Delta(T_2)$ arbitrarily close to $\hat{\pi}(t_1)$ with $\hat{\pi}_1'[a_2^*(t_2) = a_2'] > 0$ for all a_2', and such that if $\tilde{t}_1$ has $\hat{\pi}_1(\tilde{t}_1) = \hat{\pi}_1'$ and $\hat{v}_1(\tilde{t}_1) = \hat{v}_1(t_1)$ then $a_1^*(\tilde{t}_1) = a_1^*(t_1)$. As observed above, M at least as efficient as $\tilde{M}$ implies that $q(a^*(t)) \geq \tilde{q}(\tilde{a}^*(t))$ for all $t \in T$. Therefore if $Q_1(\tilde{t}_1) = \tilde{Q}_1(\tilde{t}_1)$ then it must be that $q(a_1^*(\tilde{t}_1), a_2) = \tilde{q}(\tilde{a}_1^*(\tilde{t}_1), a_2)$ for all $a_2 \in A_2$, and we can make a similar statement for every $a_1 \in A_1$. Therefore, the following proposition is equivalent to proposition III.6, and consequently proposition III.5.

Proposition III.7. *If $M = (A, \hat{y})$ and a^* are as efficient as $\tilde{M}$ and $\tilde{a}^*$ for all t, then for all t_1 it must be that in equilibrium a^* of M,*

$$Q_1(t_1) = \tilde{Q}_1(t_1). \tag{3.14}$$

In what follows I prove Proposition III.6, and thus Proposition III.5.

Proof. M as efficient as $\tilde{M}$ ensures that $Q_1(t_1) = \tilde{Q}_1(t_1) = 1$ when $\tilde{a}_1^*(t_1) =$"PAY", so I focus on the case that $\tilde{a}_1^*(t_1) =$"DON'T PAY."

I split the proof into two cases (see figure 3.1).

- Case 1 applies when t_1, the type of agent one, puts probability 1 on agent two's valuation being less than one (the thick line segment indicated in figure 1). If

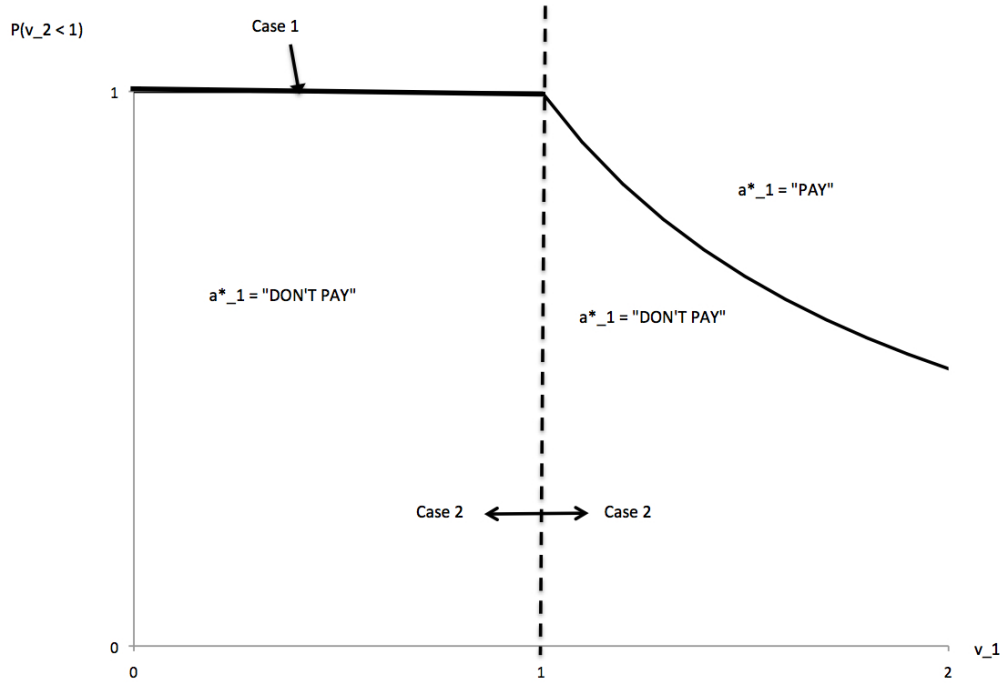


Figure 3.1: Cases 1 and 2 for the proof of Proposition III.6

$a_1^*(t_1)$ ="DON'T PAY" then it must be that $\hat{v}_1(t_1) \leq 1$. In this case I show that $Q_1(t_1) = 0 = \tilde{Q}_1(t_1)$.

- Case 2 applies when t_1 puts probability less than 1 on agent two's valuation being below one (or equivalently, puts some probability on agent two's valuation being at least one), and $\tilde{a}_1^*(t_1)$ ="DON'T PAY" (the area under the curve in figure 1). The proof of Case 2 uses Case 1, by showing that if there was an improvement in efficiency for t_1, then there would be a type of agent 1 fitting Case 1 that would also have an improvement in efficiency. Because Case 1 rules this out, it must be that there is no increase in efficiency for t_1.

The formal statement of Case 1 is as follows:

Case 1. If M and a^* is at least as efficient as $\tilde{M}$ and $\tilde{a}^*$, then for all t_1 such that

$\hat{\pi}_1(t_1)[\hat{v}_2(t_2) < 1] = 1$ and $\hat{v}_1(t_1) < 1$,

$$Q_1(t_1) = 0 = \tilde{Q}_1(t_1) \tag{3.15}$$

in equilibrium.

proof Fix some t_1' that fits case 1. The exact valuation of t_1' is unimportant, as we will show the result for all types of agent 1 with the same beliefs as t_1' and valuations less than 1. To do that, we look at the maximum expected probability of the public good being produced among those types:
Define $t'_{1,\tilde{v}}$ by $\hat{v}_1(t'_{1,\tilde{v}}) = \tilde{v}$ and $\hat{\pi}_1(t'_{1,\tilde{v}}) = \hat{\pi}_1(t_1')$ (so for example $t_1' = t'_{1,\hat{v}_1(t_1')}$). Then define

$$\overline{Q} = \max_{\tilde{v}<1} \; Q_1(t'_{1,\tilde{v}}, a_1^*(t'_{1,\tilde{v}})). \tag{3.16}$$

I shall prove the proposition by showing that $\overline{Q} = 0$. A pure strategy equilibrium and finite action space imply the maximum that defines $\overline{Q}$ exists.

The proof of Case 1 consists of two parts. I start by positing the existence of actions and types with certain properties that will be useful for the proof. Lemma III.8 describes these posited types and actions, and then shows that the requirement that M be as efficient as $\tilde{M}$ everywhere, combined with incentive compatibility for these and other types, implies that $\overline{Q} = 0$. Lemma III.9 then proves that actions and types matching those posited in Lemma III.8 exist.

Lemma III.8. *If there are actions* $\{\overline{a}_1, \underline{a}_1\} \in A_1$ *and* $\{\hat{a}_2, \overline{a}_2, \underline{a}_2\} \in A_2$ *and a sequence of subsets of* T *indicated by* $T^n = T_1^n \times T_2^n = \left\{\overline{t}_1^n, \underline{t}_1^n\right\} \times \left\{\hat{t}_2^n, \overline{t}_2^n, \underline{t}_2^n\right\}$ *such that*

1. *for all* n*, the following are true of* $\overline{a}_1$ *and* $\overline{t}_1^n$*:*

   - $a_1^*(\overline{t}_1^n) = \overline{a}_1$
   - $\hat{v}_1(\overline{t}_1^n) > 1$*; furthermore* $\tilde{a}_1^*(\overline{t}_1^n) =$ *"PAY"*

- *Given a valuation of* $v_1 = 1$, $\overline{a}_1$ *is a best response to* $\hat{\pi}_1(t_1')$*; within that set of best responses* $\overline{a}_1$ *is a best response to* $\hat{a}_2$*; within that set* $\overline{a}_1$ *is a best response to* $\underline{a}_2$.

2. *for all* n*, the following are true of* $\underline{a}_1$ *and* $\underline{t}_1^n$*:*

   - $a_1^*(\underline{t}_1^n) = \underline{a}_1$
   - $\hat{v}_1(\underline{t}_1^n) < 1$
   - *Given a valuation of* $v_1 = 1$, $\underline{a}_1$ *is a best response to* $\hat{\pi}_1(t_1')$*; within that set of best responses* $\underline{a}_1$ *is a best response to* $\hat{a}_2$*; within that set* $\underline{a}_1$ *is a best response to* $\underline{a}_2$.

3. *for all* n*, the following are true of* $\hat{a}_2$ *and* $\hat{t}_2^n$*:*

   - $a_2^*(\hat{t}_2^n) = \hat{a}_2$
   - $\hat{v}_2(\hat{t}_2^n) < 1$ *and* $\lim_{n\to\infty} \hat{v}_2(\hat{t}_2^n) = 1$
   - *Given* $\hat{v}_2(\hat{t}_2^n)$, $\hat{a}_2$ *is a best response to* $\overline{a}_1$*; within that set of best responses* $\hat{a}_2$ *is a best response to* $\underline{a}_1$.

4. *for all* n*, the following are true of* $\overline{a}_2$ *and* $\overline{t}_2^n$*:*

   - $a_2^*(\overline{t}_2^n) = \overline{a}_2$
   - $\hat{v}_2(\overline{t}_2^n) > 1$ *and* $\lim_{n\to\infty} \hat{v}_2(\overline{t}_2^n) = 1$
   - *Given* $\hat{v}_2(\overline{t}_2^n)$, $\overline{a}_2$ *is a best response to* $\overline{a}_1$*; within that set of best responses* $\overline{a}_2$ *is a best response to* $\underline{a}_1$.

5. *for all* n*, the following are true of* $\underline{a}_2$ *and* $\underline{t}_2^n$*:*

   - $a_2^*(\underline{t}_2^n) = \underline{a}_2$
   - $0 < \hat{v}_2(\underline{t}_2^n) < 1 - \hat{v}_1(\underline{t}_1^n)$

- *Given* $\hat{v}_2(\underline{t}_2^n)$, $\underline{a}_2$ *is a best response to* $\overline{a}_1$*; within that set of best responses* $\underline{a}_2$ *is a best response to* $\underline{a}_1$.

then $\overline{Q} = 0$.

Proof. Define $\hat{Q} \equiv \hat{\pi}_1(t_1')[q(\underline{a}_1, a_2^*(t_2)) = 1]$, the expected probability that the public good will be produced when the action $\underline{a}_1$ is played and given the beliefs of agent t_1' . By monotonicity of $Q_1(t_1)$ when beliefs are fixed and valuations are varied (implied by incentive compatibility) and the fact that $\underline{a}_1$ is a best response to $\hat{\pi}_1(t_1')$ for valuation 1,

$$\hat{Q} \geq \overline{Q}. \tag{3.17}$$

Thus $\underline{a}_1$ will fulfill the role of the action played by agents with valuation less than 1 and beliefs that are (roughly) the same as t_1'. For simplicitly, assume $\hat{Q} = \overline{Q}$. This is without loss of generality because I will show that $\hat{Q} = 0$, which through equation (3.17) implies $\overline{Q} = 0$. Therefore to avoid extra notation I assume that $\hat{Q}$ and $\overline{Q}$ are equal.

The proof of lemma III.8 uses incentive compatibility constraints, and the requirement that M be as efficient as $\tilde{M}$ everywhere, to pin down the outcomes when the posited actions are played. The proof has the following structure. Sublemma 1 uses the assumed best response properties of $\overline{a}_1$ and $\underline{a}_1$ to establish facts about the relative transfers and production of the public good when agent one plays $\overline{a}_1$ and $\underline{a}_1$. Sublemma 2 uses those facts and the descriptions of the posited actions and types to further describe the outcomes of these actions under M. Sublemma 3 then uses the results of sublemmas 1 and 2 to show that if M is incentive compatible and everywhere as efficient as $\tilde{M}$, then $\overline{Q} = 0$.

The following sublemma uses the best response properties assumed of $\overline{a}_1$ and $\underline{a}_1$ to show that the differences in outcomes between $\overline{a}_1$ and $\underline{a}_1$ must satisfy certain conditions.

Sublemma 1. The outcomes for actions $\overline{a}_1$ and $\underline{a}_1$ obey the following conditions:

1. Given beliefs $\hat{\pi}_1(t_1')$ the following holds:

$$E_{\hat{\pi}_1(t_1')}[\tau_1(\overline{a}_1, a_2^*(t_2)) - \tau_1(\underline{a}_1, a_2^*(t_2))] = E_{\hat{\pi}_1(t_1')}[q(\overline{a}_1, a_2^*(t_2)) - q(\underline{a}_1, a_2^*(t_2))] \quad (3.18)$$

2. When $a_2 = \hat{a}_2$, the following holds:

$$\tau_1(\overline{a}_1, \hat{a}_2) - \tau_1(\underline{a}_1, \hat{a}_2) = q(\overline{a}_1, \hat{a}_2) - q(\underline{a}_1, \hat{a}_2) \quad (3.19)$$

3. When $a_2 = \underline{a}_2$, the following holds:

$$\tau_1(\overline{a}_1, \underline{a}_2) - \tau_1(\underline{a}_1, \underline{a}_2) = q(\overline{a}_1, \underline{a}_2) - q(\underline{a}_1, \underline{a}_2) \quad (3.20)$$

Proof. The sublemma follows from the assumptions that if $v_1 = 1$ then $\overline{a}_1$ and $\underline{a}_1$ are best responses to $\hat{\pi}_1(t_1')$, which implies equation (3.18); and that within that set of best responses both actions are best responses to $\hat{a}_2$, which implies equation (3.19); and that within that set both actions are best responses to $\underline{a}_2$, which implies equation (3.20).

□

Now I show the outcomes when $a \in \{\overline{a}_1, \underline{a}_1\} \times \{\hat{a}_2, \overline{a}_2, \underline{a}_2\}$. Define the following value, which corresponds to the lowest transfer agent 2 can pay when agent 1 plays $\overline{a}_1$:

$$\underline{\tau}_2(\overline{a}_1) \equiv \min_{a_2' \in A_2} \tau_2(\overline{a}_1, a_2'). \quad (3.21)$$

Sublemma 2 shows that the properties stated in lemma III.8, combined with sublemma 1, determine the following features of the outcomes under M when $a \in \{\overline{a}_1, \underline{a}_1\} \times \{\hat{a}_2, \overline{a}_2, \underline{a}_2\}$:

$q(a_1, a_2)$			
Actions	$\underline{a}_2$	$\hat{a}_2$	$\overline{a}_2$
$\overline{a}_1$	1	1	1
$\underline{a}_1$	0	0	1

$\tau_2(a_1, a_2)$			
Actions	$\underline{a}_2$	$\hat{a}_2$	$\overline{a}_2$
$\overline{a}_1$	$\underline{\tau}_2(\overline{a}_1)$	$\underline{\tau}_2(\overline{a}_1)$	$\underline{\tau}_2(\overline{a}_1)$
$\underline{a}_1$	$\underline{\tau}_2(\overline{a}_1)$	$\underline{\tau}_2(\overline{a}_1)$	$\underline{\tau}_2(\overline{a}_1) + 1$

Table 3.2: Outcomes under M implied by sublemmas 1 and 2

(Because we have strict budget balance, $\tau_1(a)$ is also determined.) These values will be useful in showing that the type t_1^* (which will be constructed in sublemma 3) prefers $\underline{a}_1$ to $\overline{a}_1$. Sublemma 2 establishes the above tables through a series of claims.

Sublemma 2. Given the conditions specified in lemma III.8, the following claims are true of $\hat{y}(a)$ when $a \in \{\overline{a}_1, \underline{a}_1\} \times \{\hat{a}_2, \overline{a}_2, \underline{a}_2\}$.

1. For all $a_2 \in \{\hat{a}_2, \overline{a}_2, \underline{a}_2\}$,

$$q(\overline{a}_1, a_2) = 1 \tag{3.22}$$

and

$$\tau_2(\overline{a}_1, a_2) = \underline{\tau}_2(\overline{a}_1). \tag{3.23}$$

2. For $a_2 \in \{\hat{a}_2, \underline{a}_2\}$,

$$q(\underline{a}_1, a_2) = 0 \tag{3.24}$$

and

$$\tau_2(\underline{a}_1, a_2) = \underline{\tau}_2(\overline{a}_1) \tag{3.25}$$

3. For $\overline{a}_2$, the following are true:

$$q(\underline{a}_1, \overline{a}_2) = 1 \tag{3.26}$$

and

$$\tau_2(\underline{a}_1, \overline{a}_2) = \underline{\tau}_2(\overline{a}_1) + 1 \tag{3.27}$$

Proof. I show each claim in turn. Some claims require the earlier claims for their proof.

Claim 1: To prove equation (3.22), recall that by assumption $\tilde{a}_1^*(\overline{t}_1^n) =$"PAY ALL" and $\hat{v}_1(\overline{t}_1^n) > 1$ for all n. Therefore given that M is at least as efficient as $\tilde{M}$, for all $a_2 \in A_2$,

$$q(\overline{a}_1, a_2) = 1 \tag{3.28}$$

which is equation (3.22).

Equation (3.23) is an immediate implication of equations (3.21) and (3.22) and the focus lemma.

Claim 2:

Equation (3.25) is a straightforward implication of claim 1 and sublemma 1 conditions 2 and 3. Substituting equations (3.22) and (3.23) into equation (3.20) yields:

$$(1 - \underline{\tau}_2(\overline{t}_1)) - \tau_1(\underline{a}_1, \hat{a}_2) = 1 - q(\underline{a}_1, \hat{a}_2)$$

which implies

$$q(\underline{a}_1, \hat{a}_2) - \tau_1(\underline{a}_1, \hat{a}_2) = \underline{\tau}_2(\overline{t}_1).$$

By budget balance, $\tau_2(\underline{a}_1, \hat{a}_2) = q(\underline{a}_1, \hat{a}_2) - \tau_1(\underline{a}_1, \hat{a}_2)$. Substituting, we get

$$\tau_2(\underline{a}_1, \hat{a}_2) = \underline{\tau}_2(\overline{t}_1). \tag{3.29}$$

A parallel argument using equations (3.19), (3.22) and (3.23) shows that

$$\tau_2(\underline{a}_1, \underline{a}_2) = \underline{\tau}_2(\overline{t}_1). \tag{3.30}$$

Then equations (3.29) and (3.30) together imply equation (3.25).

Now I prove equation (3.24). By assumption on $\underline{t}_2^n$, for any n we have $\hat{v}_1(\underline{t}_1^n) + \hat{v}_2(\underline{t}_2^n) < 1$. Then M as efficient as $\tilde{M}$ implies that

$$q(\underline{a}_1, \underline{a}_2) = 0. \tag{3.31}$$

Claim 1 and equations (3.25) and (3.31) imply the following transfers and outcomes regarding the public good being produced depending on the actions (the "?" indicates the value we want to determine):

$q(a_1, a_2) =$		
Actions	$\underline{a}_2$	$\hat{a}_2$
$\overline{a}_1$	1	1
$\underline{a}_1$	0	?

$\tau_2(a_1, a_2) =$		
Actions	$\underline{a}_2$	$\hat{a}_2$
$\overline{a}_1$	$\underline{\tau}_2(\overline{a}_1)$	$\underline{\tau}_2(\overline{a}_1)$
$\underline{a}_1$	$\underline{\tau}_2(\overline{a}_1)$	$\underline{\tau}_2(\overline{a}_1)$

Table 3.3: Outcomes under M implied by claim 1 and equations (3.25) and (3.31)

Given the construction of $\underline{t}_2^n$ it is immediate that if $q(\underline{a}_1, \hat{a}_2) = 1$ then $\underline{t}_2^n$ would want to deviate to $\hat{a}_2$ for all n. Therefore

$$q(\underline{a}_1, \hat{a}_2) = 0. \tag{3.32}$$

Together (3.31) and (3.32) imply equation (3.24).

Claim 3:

Given that $\hat{v}_2(\overline{t}_2^n) > 1$, equation (3.11) and M as efficient as $\tilde{M}$ implies that

$$q(a_1, \overline{a}_2) = 1 \quad \forall a_1 \in A_1$$

which implies in particular that $q(\underline{a}_1, \overline{a}_2) = 1$, which proves equation (3.26).

To determine $\tau_2(\underline{a}_1, \overline{a}_2)$, I observe that claims 1 and 2 with equation (3.26) establish the following values for transfers and whether the public good is produced (the "?" indicates the value we want to determine):

$q(a_1, a_2) =$			
Actions	$\underline{a}_2$	$\hat{a}_2$	$\overline{a}_2$
$\overline{a}_1$	1	1	1
$\underline{a}_1$	0	0	1

$\tau_2(a_1, a_2) =$			
Actions	$\underline{a}_2$	$\hat{a}_2$	$\overline{a}_2$
$\overline{a}_1$	$\underline{\tau}_2(\overline{a}_1)$	$\underline{\tau}_2(\overline{a}_1)$	$\underline{\tau}_2(\overline{a}_1)$
$\underline{a}_1$	$\underline{\tau}_2(\overline{a}_1)$	$\underline{\tau}_2(\overline{a}_1)$	?

Table 3.4: Outcomes under M implied by claims 1 and 2 with equation (3.26)

Equation (3.27) is then implied by incentive compatibility. If

$$\tau_2(\underline{a}_1, \overline{a}_2) < \underline{\tau}_2(\overline{a}_1) + 1$$

then $\hat{v}_2(\hat{t}_2^n) \to_- 1$ implies that for some n, $\hat{t}_2^n$ would want to deviate from $\hat{a}_2$ to $\overline{a}_2$. However, if

$$\tau_2(\underline{a}_1, \overline{a}_2) > \underline{\tau}_2(\overline{a}_1) + 1$$

then $\hat{v}_2(\overline{t}_2^n) \to_+ 1$ implies that for some n, $\overline{t}_2^n$ would want to deviate to $\hat{a}_2$. Therefore equation (3.27) must hold.

□

The following sublemma uses the results from sublemmas 1 and 2 to prove lemma III.8. I construct a new type, t_1^* for agent 1 that under the equilibrium of the all or nothing mechanism always achieves an efficient outcome. Among actions under M that always produce the public good (and therefore achieve an efficient outcome), t_1^*'s optimal action is $\overline{a}_1$. However, if $\overline{Q} > 0$ then t_1^* prefers $\underline{a}_1$ to $\overline{a}_1$, which implies that $a_1^*(t_1^*)$ must not always produce the public good and therefore does not always

achieve an outcome as efficient as under $\tilde{M}$. Therefore if M is as efficient as $\tilde{M}$ then it must be that $\overline{Q} = 0$.

Sublemma 3. If the conditions of sublemma 1 hold and if $\overline{Q} > 0$ then there exists a t_1^* such that

$$\tilde{a}_1^*(t_1^*) = \text{“}PAY\text{”} \tag{3.33}$$

and

$$Q_1(t_1^*, a_1^*(t_1^*)) < 1. \tag{3.34}$$

Proof. Recall that I have assumed without loss of generality that

$$\hat{Q} \equiv \hat{\pi}_1(t_1')[q(\underline{a}_1, a_2^*(t_2)) = 1] = \overline{Q}. \tag{3.35}$$

I now define a type $\overline{t}_1^*$ that will help me construct t_1^*. Let the type $\overline{t}_1^*$ be defined by, for some n, $\hat{v}_1(\overline{t}_1^*) = \hat{v}_1(\overline{t}_1^n)$ and for some ϵ_*,

$$\hat{\pi}_1(\overline{t}_1^*)[E] \equiv \hat{\pi}_1(t_1')[E] \cdot (1 - \epsilon_*) + (\epsilon_* - \epsilon_*^2) \cdot I_{\hat{t}_2^n \in E}$$

$$+(\epsilon_*^2 - \epsilon_*^3) \cdot I_{\underline{t}_2^n \in E} + \epsilon_*^3 \cdot I_{\overline{t}_2^n \in E} \tag{3.36}$$

where ϵ_* is small enough that the focus lemma implies that $a_1^*(\overline{t}_1^*)$ is a best response, conditional on $\hat{v}_1(\overline{t}_1^*)$, to $\hat{\pi}_1(t_1')$; within that set of best responses $a_1^*(\overline{t}_1^*)$ is a best response to $\hat{a}_2$; within that set $a_1^*(\overline{t}_1^*)$ is a best response to $\underline{a}_2$. Furthermore I require that $\hat{\pi}_1(\overline{t}_1^*)[\hat{v}_2(t_2) < 1] = 1 - \epsilon_*^3 > \frac{1}{\hat{v}_1(\overline{t}_1^n)}$ to ensure that $\tilde{a}_1^*(\overline{t}_1^n) =$“PAY”.

This best response condition result implies that $U_1(\overline{t}_1^*, a_1^*(\overline{t}_1^*)) = U_1(\overline{t}_1^*, \overline{a}_1)$ by construction of $\overline{a}_1$, as $\overline{a}_1$ satisfies the same best response condition for all $v_1 \geq 1$, and therefore in particular $\hat{v}_1(\overline{t}_1^n)$. Furthermore, by M as efficient as $\tilde{M}$ both actions pro-

duce the public good whatever action the other agent takes. Therefore without loss of generality we can assume that $\overline{a}_1 = a_1^*(\overline{t}_1^*)$.

Now I look at types with the same beliefs as $\overline{t}_1^*$ but with different valuations. One such type will be our t_1^*. For any v, define $t_1^*(v)$ by $\hat{v}_1(t_1^*(v)) = v$ and $\hat{\pi}_1(t_1^*(v)) = \hat{\pi}_1(\overline{t}_1^*))$. I first calculate the difference in expected utility for such types when taking the action $\overline{a}_1$ compared to the action $\underline{a}_1$. Equation (3.36) translated into relative utilities is as follows:

$$\begin{aligned} U_1(t_1^*(v), \overline{a}_1) - U_1(t_1^*(v), \underline{a}_1) = (1-\epsilon_*) \cdot E_{\hat{\pi}_1(t_1')}[u_1(v, \hat{y}(\overline{a}_1, a_2^*(t_2))) - u_1(v, \hat{y}((\underline{a}_1, a_2^*(t_2)))] \\ +(\epsilon_* - \epsilon_*^2) \cdot (u_1(v, \hat{y}(\overline{a}_1, \hat{a}_2)) - u_1(v, \hat{y}(\underline{a}_1, \hat{a}_2))) \\ +(\epsilon_*^2 - \epsilon_*^3) \cdot (u_1(v, \hat{y}(\overline{a}_1, \underline{a}_2)) - u_1(v, \hat{y}(\underline{a}_1, \underline{a}_2))) \\ +\epsilon_*^3 \cdot (u_1(v, \hat{y}(\overline{a}_1, \overline{a}_2)) - u_1(v, \hat{y}(\underline{a}_1, \overline{a}_2))) \end{aligned} \tag{3.37}$$

Sublemma 1, condition 1 implies that for any v, we can make the following substitution for the first term on the right side of equation (3.37):

$$E_{\hat{\pi}_1(t_1')}[u_1(v, \hat{y}(\overline{a}_1, a_2^*(t_2))) - u_1(v, \hat{y}(\underline{a}_1, a_2^*(t_2))] = (1 - \overline{Q}) \cdot (v - 1). \tag{3.38}$$

To find expressions for the other terms in equation (3.37), note that for any v_1 and $a_2 \in A_2$,

$$u_1(v, \hat{y}(\overline{a}_1, a_2)) - u_1(v, \hat{y}(\underline{a}_1, a_2)) = [q(\overline{a}_1, a_2) - q(\underline{a}_1, a_2)] \cdot v - (\tau_1(\overline{a}_1, a_2) - \tau_1(\underline{a}_1, a_2)) \tag{3.39}$$

that is, the difference in utility is the difference in whether the public good is produced

times the valuation, minus the difference in transfers. We can determine values for equation (3.39) when $a_2 \in \{\hat{a}_2, \underline{a}_2, \overline{a}_2\}$ using the results for sublemma 2. We can write the following tables for transfers and production of the public good based on certain specific actions (note that for this table I list transfers by agent 1, not agent 2 as in previous tables):

$q(a_1, a_2) =$			
Actions	$\underline{a}_2$	$\hat{a}_2$	$\overline{a}_2$
$\overline{a}_1$	1	1	1
$\underline{a}_1$	0	0	1

$\tau_1(a_1, a_2) =$			
Actions	$\underline{a}_2$	$\hat{a}_2$	$\overline{a}_2$
$\overline{a}_1$	$1 - \underline{\tau}_2(\overline{a}_1)$	$1 - \underline{\tau}_2(\overline{a}_1)$	$1 - \underline{\tau}_2(\overline{a}_1)$
$\underline{a}_1$	$-\underline{\tau}_2(\overline{a}_1)$	$-\underline{\tau}_2(\overline{a}_1)$	$-\underline{\tau}_2(\overline{a}_1)$

Table 3.5: Outcomes under M

These tables gives us the information needed to use equation (3.39) to assign values to the last three terms in equation (3.37). I make these substitutions, and substituting equation (3.38), into equation (3.37) to get

$$U_1(t_1^*(v), \overline{a}_1) - U_1(t_1^*(v), \underline{a}_1) = (1 - \epsilon_*) \cdot \left[(1 - \overline{Q}) \cdot (v - 1)\right]$$

$$+(\epsilon_* - \epsilon_*^2) \cdot [v - 1] + (\epsilon_*^2 - \epsilon_*^3) \cdot [v - 1] + \epsilon_*^3 \cdot [-1]$$

$$= \left[(1 - \epsilon_*) \cdot (1 - \overline{Q}) + (\epsilon_* - \epsilon_*^3)\right] \cdot (v - 1) - \epsilon_*^3 \tag{3.40}$$

By comparison, the similar result for the actions "PAY" and "DON'T PAY" under $\tilde{M}$ is

$$U_1^{\tilde{M}}(t_1^*(v), \text{“}PAY\text{”}) - U_1^{\tilde{M}}(t_1^*(v), \text{“}DON'TPAY\text{”}) = (1 - \epsilon_*^3) \cdot (v - 1) - \epsilon_*^3 \tag{3.41}$$

This is easily verified by inspection of the equilibrium $\tilde{a}^*$ of the All or Nothing Mechanism $\tilde{M}$.

Note that by inspection equation (3.41) is greater than (3.40) for $v > 1$ given that $\overline{Q} > 0$. I now generate a type where equation (3.41) is equal to 0, which implies that the agent plays "PAY" under $\tilde{M}$. At the same time that implies that equation (3.40) is negative, although I calculate this explicitly. I show these two facts and then use them to prove $Q_1(t_1^*) < 1$.

Define $t_1^* = t_1^*(\frac{1}{1-\epsilon_*^3})$. By construction of $\hat{\pi}_1(\overline{t}_1^*)$ it is the case that $\hat{\pi}_1(t_1^*)[\hat{v}_2(t_2) < 1] = 1 - \epsilon_*^3 = \frac{1}{\hat{v}_1(t_1^*)}$. That implies that $\tilde{a}_1^*(t_1^*) =$ "PAY", satisfying equation (3.33).

To prove equation (3.34), I observe that by construction equation (3.40) is equal to 0 when $v = \hat{v}_1(t_1^*)$. Then by inspection equation (3.41) is strictly negative given that $\overline{Q} > 0$ by assumption. So $U_1(t_1^*, \underline{a}_1) > U_1(t_1^*, \overline{a}_1)$.

Because $\overline{a}_1$ is a best response for $\overline{t}_1$, it must be that among actions that produce the public good with certainty $\overline{a}_1$ is transfer-minimizing given $\hat{\pi}_1(\overline{t}_1^*)$ and therefore optimal for t_1^* within that set. Therefore if $U_1(t_1^*, \underline{a}_1) > U_1(t_1^*, \overline{a}_1)$ then it must be that $a_1^*(t_1^*)$ does not always produce the public good, and in fact that $Q_1(t_1^*, a_1^*(t_1^*)) < 1$. That proves equation (3.34) and completes the proof of the sublemma. □

Sublemma 3 implies that M as efficient as $\tilde{M}$ requires that $\overline{Q} = 0$, and by definition of $\overline{Q}$ it must be that $Q_1(t_1') \leq \overline{Q} = 0$, which proves equation (3.15). This completes the proof of lemma III.8. □

Lemma III.9. *There exists actions $\{\overline{a}_1, \underline{a}_1\} \in A_1$ and $\{\hat{a}_2, \overline{a}_2, \underline{a}_2\} \in A_2$ and a sequence of subsets of T indicated by $T^n = T_1^n \times T_2^n = \left\{\overline{t}_1^n, \underline{t}_1^n\right\} \times \left\{\hat{t}_2^n, \overline{t}_2^n, \underline{t}_2^n\right\}$ that satisfy the conditions in lemma III.8.*

Proof. I start by constructing the type spaces $\{T^n\}$, and then show the existence of actions fitting the conditions on $\{\overline{a}_1, \underline{a}_1\}$ and $\{\hat{a}_2, \overline{a}_2, \underline{a}_2\}$.

Each T^n will belong to a class of subsets of the universal type space of the form $T^{n,m,\epsilon}$, defined in the following way: For n and m positive integers and $\epsilon > 0$, let

$T^{n,m,\epsilon} = T_1^{n,m,\epsilon} \times T_2^{n,m,\epsilon}$ where $T_1^{n,m,\epsilon} = \{\overline{t}_1, \underline{t}_1, t_1^0\}$ and $T_2^{n,m,\epsilon} = \{\hat{t}_2, \overline{t}_2, \underline{t}_2\}$. The valuations and beliefs of the types are defined below. First I comment briefly on the role of n, m and ϵ respectively.

- n. Taking the limit as n goes to infinity will ensure that the best response conditions for $\overline{a}_1$ and $\underline{a}_1$ hold, by use of the focus lemma. It will also ensure that the conditions on the valuations of $\hat{t}_2$ and $\overline{t}_2$ hold. Taking a sequence of n (and appropriately chosen m and ϵ) will give us a sequence of $T^{n,m,\epsilon}$ from which the sequence T^n will be drawn.

- m. For each n, taking the limit as m goes to infinity will ensure that the best response condition for $\overline{a}_1$ and $\underline{a}_1$ hold specifically at the valuation $v_1 = 1$. m positive ensures that the condition on the valuations of $\overline{t}_1^n$, $\underline{t}_1^n$ and $\underline{t}_2^n$ hold. m will be chosen as a function of n.

- ϵ. Choosing a sufficiently small ϵ for each n and m combination will ensure that the best response conditions for $\hat{a}_2, \underline{a}_2$ and $\overline{a}_2$ hold.

Valuations: For simplicity I present the valuations in table form.

$\hat{v}_1(t_1)$	
$\overline{t}_1$	$1 + \frac{1}{m}$
$\underline{t}_1$	$1 - \frac{1}{m}$
t_1^0	1

$\hat{v}_2(t_2)$	
$\underline{t}_2$	$\frac{1}{2 \cdot m}$
$\hat{t}_2$	$1 - \frac{1}{n}$
$\overline{t}_2$	$1 + \frac{1}{n}$

Table 3.6: Valuations of the constructed types in $T^{n,m,\epsilon}$

Note that if $m > 0$ (as we require), and we let $n \to \infty$ then the valuations of these types match the requirements on the corresponding types $\overline{t}_1^n, \underline{t}_1^n, \underline{t}_2^n, \hat{t}_2^n$ and $\overline{t}_2^n$. (t_1^0 has no corresponding type in T^n.)

Beliefs: the beliefs of the types of agent 1 are a function of n, while the beliefs of the types of agent 2 are a function of ϵ. For all $t_1 \in T_1^{n,m,\epsilon}$ beliefs are defined by the following $p_2 \in \Delta T_2$. For all $E \subset T_2$,

$$p_2[E] = \hat{\pi}_1(\bar{t}_1)[E] = \hat{\pi}_1(\underline{t}_1)[E] = \hat{\pi}_1(t_1^0)[E]$$

$$= (1 - \frac{1}{n})\hat{\pi}_1(t_1')[E] + \frac{1}{n}(1 - \frac{1}{n}) \cdot I_{\hat{t}_2 \in E} + \frac{1}{n}^2 \cdot I_{\underline{t}_2 \in E} \tag{3.42}$$

that is, types of agent 1 expect agent 2's type to be distributed according to $\hat{\pi}_1(t_1')$ with probability $1 - \frac{1}{n}$, and the other $\frac{1}{n}$ probability is assigned to $\hat{t}_2$ and $\underline{t}_2$ at a ratio of $1 - \frac{1}{n}$ to $\frac{1}{n}$ respectively.

Similarly, for all $t_2 \in T_2^{n,m,\epsilon}$ beliefs are defined by the following $p_1 \in \Delta T_1$. For all $E \subset T_1$,

$$p_1[E] = \hat{\pi}_2(\hat{t}_2)[E] = \hat{\pi}_2(\underline{t}_2)[E] = \hat{\pi}_2(\bar{t}_2)[E]$$

$$= (1 - \epsilon) \cdot I_{\bar{t}_1 \in E} + \epsilon \cdot I_{\underline{t}_1 \in E} \tag{3.43}$$

that is, types of agent 2 expect agent 1's type to be $\bar{t}_1$ with probability $1 - \epsilon$, and expect that agent 1's type will be $\underline{t}_1$ otherwise.

I now construct a series of $T^{n,m,\epsilon}$ by first holding n, m constant and finding an appropriate ϵ for each (n, m) combination. Then I hold n constant and find an appropriate m for each n. Finally I take the sequence created by varying n and find a subsequence of that sequence that I use to construct $\{T^n\}$.

Holding n and m fixed. For each possible n, m the focus lemma implies that we can find an $\epsilon(n, m)$ small enough that $\hat{t}_2$, $\underline{t}_2$ and $\bar{t}_2$ best respond (given their valuations) to $\bar{t}_1$, and conditional on that best respond to $\underline{t}_1$. Fix such an $\epsilon(n, m)$ for all n and m.

Holding n fixed. Consider the set $\{T^{n,m,\epsilon(n,m)}\}_{m\in\{1,2,...\}}$. Define

$$a^{n,m} = \left\{a_1^*(\bar{t}_1), a_1^*(\underline{t}_1), a_1^*(t_1^0), a_2^*(\underline{t}_2), a_2^*(\hat{t}_2), a_2^*(\bar{t}_2)\right\} \equiv \{\overline{a}_1^{n,m}, \underline{a}_1^{n,m}, a_1^{0,n,m}, \underline{a}_2^{n,m}, \hat{a}_2^{n,m}, \overline{a}_2^{n,m}\}. \tag{3.44}$$

where $\{\bar{t}_1, \underline{t}_1, t_1^0, \underline{t}_2, \hat{t}_2, \bar{t}_2\}$ correspond to the elements in $T_1^{n,m,\epsilon(n,m)}$ and $T_2^{n,m,\epsilon(n,m)}$.

Because M has finite actions, and we assume a pure strategy equilibrium, the sequence $\{a^{n,m}\}_{m\in\{1,2,...\}}$ has a constant subsequence, with constant term I will call a^n. Let $m(n)$ correspond to some m such that $a^{n,m(n)} = a^n$. Let $\epsilon(n) \equiv \epsilon(n, m(n))$, and let $\tilde{T}^n = T^{n,m(n),\epsilon(n,m)}$.

Sublemma 4. For the type $t_1^0 \in \tilde{T}_1^n$, the actions $a_1^*(t_1^0), a_1^*(\bar{t}_1)$ and $a_1^*(\underline{t}_1)$ are all best responses in equilibrium.

Proof. By construction, there exists arbitrarily large m' such that

$$\left\{a_1^*(\bar{t}_1), a_1^*(\underline{t}_1), a_1^*(t_1^0), a_2^*(\underline{t}_2), a_2^*(\hat{t}_2), a_2^*(\bar{t}_2)\right\} = a^n$$

for $\bar{t}_1, \underline{t}_1 \in T_1^{n,m',\epsilon(n,m')}$. That implies that for the beliefs p_2 (defined relative to $T^{n,m',\epsilon(n,m')}$) and valuations $1 + \frac{1}{m'}$ arbitrarily close to 1, $a_1^*(\bar{t}_1)$ is a best response. Similarly, given beliefs p_2 and valuations $1 - \frac{1}{m'}$ arbitrarily close to 1, $a_1^*(\underline{t}_1)$ is a best response. These facts imply that both actions are best responses for a type with beliefs p_2 and valuation 1, which describes t_1^0. □

Varying n. Now we can find a constant subsequence of the series $\{a^n\}$. Let $\overline{a}^n = \{\overline{a}_1, \underline{a}_1, a_1^0, \underline{a}_2, \hat{a}_2, \overline{a}_2\}$ be the constant term in the subsequence. Take the corresponding subsequence $\{\tilde{T}^n\}$ and let $\{T^n\}$ be defined by $T_1^n = \tilde{T}_1^n \setminus \{t_1^0\}$ and $T_2^n = \tilde{T}_2^n$. Unsurprisingly, these will be our candidates for the objects assumed in lemma III.8.

Sublemma 5. The actions $\{\overline{a}_1, \underline{a}_1\} \in A_1$ and $\{\hat{a}_2, \overline{a}_2, \underline{a}_2\} \in A_2$ and the sequence

$T^n = T_1^n \times T_2^n = \left\{\overline{t}_1^n, \underline{t}_1^n\right\} \times \left\{\hat{t}_2^n, \overline{t}_2^n, \underline{t}_2^n\right\}$ satisfy the conditions assumed in lemma III.8.

Proof. There are three types of properties posited in lemma III.8: equilibrium actions, valuations, and best response properties. I show each in turn.

Equilibrium actions: Because the types in T^n correspond to the types in $\tilde{T}^n$, where $a^n = \overline{a}^n$,

$$\left\{a_1^*(\overline{t}_1^n), a_1^*(\underline{t}_1^n), a_2^*(\underline{t}_2^n), a_2^*(\hat{t}_2^n), a_2^*(\overline{t}_2^n)\right\} = \{\overline{a}_1, \underline{a}_1, \underline{a}_2, \hat{a}_2, \overline{a}_2\}$$

which establishes the equilibrium action conditions.

Valuations: By construction, $(\hat{v}_1(\overline{t}_1^n), \hat{v}_1(\underline{t}_1^n), \hat{v}_1(\underline{t}_2^n), \hat{v}_2(\hat{t}_2^n), \hat{v}_2(\overline{t}_2^n)) = (1+\frac{1}{m(n)}, 1-\frac{1}{m(n)}, \frac{1}{2m(n)}, 1-\frac{1}{n}, 1+\frac{1}{n})$. By inspection these values satisfy the conditions assumed in lemma III.8 as $n \to_+ \infty$.

Best response properties: This is the most complicated set of properties to show. By the choice of $\epsilon(n,m)$ we have that $\underline{t}_2^n, \hat{t}_2^n$ and $\overline{t}_2^n$ all best respond to $\overline{a}_1$ given their valuations, and conditional on that best respond to $\underline{a}_1$. We also have, by construction, that their equilibrium actions are $\underline{a}_2, \hat{a}_2$, and $\overline{a}_2$ respectively, so the property is proved for these actions.

The argument for $\overline{a}_1$ and $\underline{a}_1$ is more complicated, and uses sublemma 4. Let $t_1^{0,n}$ correspond to $t_1^0 \in \tilde{T}^n$. Then for all n, sublemma 4 implies that for $t_1^{0,n}$, the actions $\overline{a}_1$ and $\underline{a}_1$ are best responses. Note that for all n, $\hat{v}_1(t_1^{0,n}) = 1$ by construction. Furthermore, $t_1^{0,n}$'s beliefs in equilibrium correspond to equation (3.42) for n tending to ∞. Therefore we can find a sufficiently large n' such that $t_1^{0,n}$'s beliefs and the focus lemma imply that any best response, given the valuation 1, must be a best response to $\hat{\pi}_1(t_1')$, and conditional on that a best response to $\hat{t}_2$, and conditional on that a best response to $\underline{t}_2$. As both $\overline{a}_1$ and $\underline{a}_1$ are best response for the $t_1^{0,n}$, the best response property for those actions is proved.

□

Sublemma 5 completes the proof of lemma III.9 by showing that there exist $\{T^n\}$ and $\{\overline{a}_1, \underline{a}_1, \hat{a}_2, \underline{a}_2, \overline{a}_2\}$ that satisfy the conditions posited in lemma III.8. □

Together, lemmas III.8 and III.9 prove Case 1. □

Case 2. If M is at least as efficient at $\tilde{M}$ then for all t_1 such that $a_1^*(t_1) =$"DON'T PAY", $\hat{\pi}_1(t_1)[\hat{v}_2(t_2) < 1] < 1$,

$$Q_1(t_1) = \tilde{Q}_1(t_1). \tag{3.45}$$

Proof. Let t_1' be a type of agent 1 that fits the conditions of case 2. Note that M as efficient as $\tilde{M}$ implies

$$Q_1(t_1') \geq \tilde{Q}_1(t_1') = \hat{\pi}_1(t_1')[\hat{v}_2(t_2) \geq 1]$$

because the public good must be produced whenever $\hat{v}_2(t_2) \geq 1$. Therefore to prove equation (3.45) only requires showing that $Q_1(t_1')$ is not strictly greater than $\tilde{Q}_1(t_1')$.

I prove case 2 (as with case 1) with two lemmas. Lemma III.10 assumes the existence of certain types related to the posited type where there is an improvement, and proves the result given the existence of those types. Lemma III.11 proves the existence of the described types.

Lemma III.10. *If there are types $\hat{t}_1$, $\overline{t}_1$ and $\underline{t}_1$ such that*

- *For $\hat{t}_1$ the following conditions hold:*

  (A1) $\hat{v}_1(\hat{t}_1) > \max\{\hat{v}_1(t_1'), 1\}$.

  (A2) $\hat{\pi}_1(\hat{t}_1)[\hat{v}_2(t_2) < 1] = \frac{1}{\hat{v}_1(\hat{t}_1)}$

(A3) $\forall a_2 \in A_2$

$$\hat{\pi}_1(\hat{t}_1)[a_2^*(t_2) = a_2,\ \hat{v}_2(t_2) < 1] \geq \hat{\pi}_1(t_1')[a_2^*(t_2) = a_2,\ \hat{v}_2(t_2) < 1]$$

(A4) $\forall a_1 \in A_1$,

$$E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1, a_2^*(t_2)) = E_{\hat{\pi}_1(t_1')}\ \tau_1(a_1, a_2^*(t_2))$$

- *For* $\bar{t}_1$ *the following conditions hold:*

  (B1) $\hat{v}_1(\bar{t}_1) = \hat{v}_1(\hat{t}_1)$

  (B2) $\hat{\pi}_1(\bar{t}_1)[\hat{v}_2(t_2) < 1] = 1$

  (B3) $\forall a_2 \in A_2$

  $$\hat{\pi}_1(\bar{t}_1)[a_2^*(t_2) = a_2,\ \hat{v}_2(t_2) < 1] \geq \hat{\pi}_1(t_1')[a_2^*(t_2) = a_2,\ \hat{v}_2(t_2) < 1]$$

  (B4) $\forall a_1 \in A_1$,

  $$E_{\hat{\pi}_1(\bar{t}_1)}\ \tau_1(a_1, a_2^*(t_2)) = E_{\hat{\pi}_1(t_1')}\ \tau_1(a_1, a_2^*(t_2))$$

- *For* $\underline{t}_1$ *the following conditions hold:*

  (C1) $\hat{v}_1(\underline{t}_1) = 0$

  (C2) $\hat{\pi}_1(\underline{t}_1) = \hat{\pi}_1(\bar{t}_1)$

then $Q_1(t_1') = \tilde{Q}_1(t_1')$.

Proof. The proof shows that if $Q_1(t_1') > \tilde{Q}_1(t_1')$ then type $\hat{t}_1$ would want to deviate to the equilibrium action of t_1'. To show this, I prove that $\hat{t}_1$ is indifferent between its own equilibrium action and the equilibrium action of $\underline{t}_1$, but unless $Q_1(t_1') = \tilde{Q}_1(t_1')$ holds, $\hat{t}_1$ prefers the equilibrium action of t_1' to the equilibrium action of $\underline{t}_1$ and therefore also to its own equilibrium action.

To show that $\hat{t}_1$ is indifferent between its own equilibrium action and the equilibrium action of $\underline{t}_1$, I will need the result that $\hat{t}_1$ is indifferent between its equilibrium action and that of $\bar{t}_1$. I prove this first in sublemma 6.

Sublemma 6. $\hat{t}_1$ is indifferent between $a_1^*(\hat{t}_1)$ and $a_1^*(\bar{t}_1)$, i.e.

$$U_1(\hat{t}_1, a_1^*(\hat{t}_1)) = U_1(\hat{t}_1, a_1^*(\bar{t}_1)) \tag{3.46}$$

Proof. Combining property (A4) of $\hat{t}_1$ and (B4) of $\bar{t}_1$ shows that for all $a_1 \in A_1$,

$$E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1, a_2^*(t_2)) = E_{\hat{\pi}_1(\bar{t}_1)}\ \tau_1(a_1, a_2^*(t_2)) \tag{3.47}$$

that is, $\hat{t}_1$ and $\bar{t}_1$ have the same expected transfer for any action.

Together (A1) and (A2) imply that $a_1^*(\hat{t}_1) =$"PAY", while (B1) and (B2) together imply that $a_1^*(\bar{t}_1) =$"PAY". Therefore, because M is as efficient as $\tilde{M}$, and all actions are played in equilibrium,

$$q(a_1^*(\hat{t}_1), a_2) = 1 = q(a_1^*(\bar{t}_1), a_2) \quad \forall a_2 \in A_2 \tag{3.48}$$

Given that $a^*(\hat{t}_1)$ and $a^*(\bar{t}_1)$ both produce the public good with certainty, and given $\hat{t}_1$ and $\bar{t}_1$ have the same expected transfer for any action, it must be that

$$E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1^*(\hat{t}_1), a_2^*(t_2)) = E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1^*(\bar{t}_1), a_2^*(t_2)) \tag{3.49}$$

or else either $\hat{t}_1$ would want to deviate to $a_1^*(\bar{t}_1)$ or $\bar{t}_1$ would want to deviate to $a_1^*(\hat{t}_1)$. Equations (3.48) and (3.49) imply equation (3.46) and complete the proof of the sublemma. □

Now I can show that $\hat{t}_1$ is indifferent between its own equilibrium action and that

of $\underline{t}_1$, by showing that $\hat{t}_1$ is indifferent between $a_1^*(\bar{t}_1)$ and $a_1^*(\underline{t}_1)$. This result will be used in the proof of sublemma 8, which shows that $\hat{t}_1$'s indifference condition only holds if $Q_1(t_1') = \tilde{Q}_1(t_1')$.

Sublemma 7. $\hat{t}_1$ is indifferent between $a_1^*(\bar{t}_1)$ and $a_1^*(\underline{t}_1)$, i.e.

$$U_1(\hat{t}_1, a_1^*(\bar{t}_1)) = U_1(\hat{t}_1, a_1^*(\underline{t}_1)) \tag{3.50}$$

Proof. The result to be shown can be written as

$$\begin{aligned} &\hat{v}_1(\hat{t}_1) \cdot Q_1(\hat{t}_1, a_1^*(\bar{t}_1)) - E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1^*(\bar{t}_1), a_2^*(t_2)) \\ &= \hat{v}_1(\hat{t}_1) \cdot Q_1(\hat{t}_1, a_1^*(\underline{t}_1)) - E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2)) \end{aligned} \tag{3.51}$$

Because $\tilde{a}_1^*(\bar{t}_1) =$"PAY", and M as efficient as $\tilde{M}$,

$$Q_1(\hat{t}_1, a_1^*(\bar{t}_1)) = 1 \tag{3.52}$$

while $\hat{v}_1(\underline{t}_1) = 0$ and M as efficient as $\tilde{M}$ implies that

$$Q_1(\hat{t}_1, a_1^*(\underline{t}_1)) = \hat{\pi}_1(\hat{t}_1)[\hat{v}_2(t_2) \geq 1]. \tag{3.53}$$

Recall (A2) requires that $\hat{\pi}_1(\hat{t}_1)[\hat{v}_2(t_2) < 1] = \frac{1}{\hat{v}_1(\hat{t}_1)}$. Therefore (A2) together with equations (3.52) and (3.53) implies that equation (3.51) can be written as

$$\hat{v}_1(\hat{t}_1) - E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1^*(\bar{t}_1), a_2^*(t_2)) = \hat{v}_1(\hat{t}_1)\left(1 - \frac{1}{\hat{v}_1(\hat{t}_1)}\right) - E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2))$$

which a little rearrangment shows is equivalent to:

$$E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1^*(\bar{t}_1), a_2^*(t_2)) - E_{\hat{\pi}_1(\hat{t}_1)}\ \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2)) = 1. \tag{3.54}$$

For any agent 1 action, condition (A4) says that $\hat{t}_1$'s expected transfer is the same as t_1''s expected transfer, while condition (B4) says that $\bar{t}_1$'s expected transfer is also the same as t_1''s expected transfer. Therefore for any action $\hat{t}_1$ and $\bar{t}_1$ have the same expected transfer. Then the expectations in equation (3.54) can be are taken with respect to $\hat{\pi}_1(\bar{t}_1)$ instead of $\hat{\pi}_1(\hat{t}_1)$:

$$E_{\hat{\pi}_1(\bar{t}_1)}\ \tau_1(a_1^*(\bar{t}_1), a_2^*(t_2)) - E_{\hat{\pi}_1(\bar{t}_1)}\ \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2)) = 1. \tag{3.55}$$

To show that (3.55) holds I use Case 1 (to show that the left side is at least as great as 1) and M as efficient as $\tilde{M}$ (to show the left side is no greater than 1).

I can use Case 1 because condition (B2) implies that for any $v_1 \in (0,1)$ and any $t_1 \in T_1$ such that $\hat{v}_1(t_1) = v_1$ and $\hat{\pi}_1(t_1) = \hat{\pi}_1(\bar{t}_1) = \hat{\pi}_1(\underline{t}_1)$, Case 1 applies to t_1. Then $Q_1(t_1) = 0 = Q_1(\underline{t}_1)$. Because t_1 and $\bar{t}_1$ evaluate expected transfers the same, it must be that t_1 is indifferent between its equilibrium action and $a_1^*(\underline{t}_1)$. The difference in utility for t_1 of playing $a_1^*(\bar{t}_1)$ and $a_1^*(\underline{t}_1)$ is v_1 minus the difference in expected transfers (because playing $a_1^*(\bar{t}_1)$ ensures the public good is produced, while t_1 expects the public good to never be produced under its equilibrium action). Therefore for t_1 to not want to deviate for any $v_1 < 1$ it must be that

$$E_{\hat{\pi}_1(\bar{t}_1)}\ \tau_1(a_1^*(\bar{t}_1), a_2^*(t_2)) - E_{\hat{\pi}_1(\bar{t}_1)}\ \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2)) \geq 1. \tag{3.56}$$

Now we use the fact that M is as efficient as $\tilde{M}$ to prove an analogous argument for $v_1 > 1$. Let t_1 be defined as in the last paragraph, except now take $v_1 \in (1, \overline{v})$. Clearly with $\hat{\pi}_1(t_1)[\hat{v}_2(t_2) < 1] = \hat{\pi}_1(\bar{t}_1)[\hat{v}_2(t_2) < 1] = 1$, $\tilde{a}_1^*(t_1) =$"PAY", and therefore M as efficient as $\tilde{M}$ implies that $Q_1(t_1) = 1$. Given that t_1 and $\bar{t}_1$ have the same beliefs, they evaluate transfers the same, and since both agents' equilibrium actions produce the public good with certainty t_1 must be indifferent between $a_1^*(t_1)$ and $a_1^*(\bar{t}_1)$. The difference in utility for t_1 of playing $a_1^*(\bar{t}_1)$ and $a_1^*(\underline{t}_1)$ is v_1 minus the

difference in expected transfers (because playing $a_1^*(\bar{t}_1)$ ensures the public good is produced, while t_1 expects the public good to never be produced under its equilibrium action). Therefore for t_1 to not want to deviate for any $v_1 > 1$ it must be that

$$E_{\hat{\pi}_1(\bar{t}_1)}\ \tau_1(a_1^*(\bar{t}_1), a_2^*(t_2)) - E_{\hat{\pi}_1(\bar{t}_1)}\ \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2)) \leq 1. \tag{3.57}$$

Clearly equations (3.56) and (3.57) imply equation (3.55). As noted above, equation (3.55) (and therefore equation (3.54) is equivalent to proving the sublemma. □

Sublemma 8. If

$$Q_1(t_1') > \tilde{Q}_1(t_1') \tag{3.58}$$

then

$$U_1(\hat{t}_1, a_1^*(t_1')) > U_1(\hat{t}_1, a_1^*(\underline{t}_1)) \tag{3.59}$$

Proof. It is notable that equation (3.58) mentions $\tilde{Q}_1(t_1')$ while equation (3.59) mentions $a_1^*(\underline{t}_1)$. The connection is the following. By construction of $\tilde{a}^*$,

$$\tilde{Q}_1(t_1') = \hat{\pi}_1(t_1')[\hat{v}_2(t_2) \geq 1]$$

while M as efficient as $\tilde{M}$ and $\hat{v}_1(\underline{t}_1) = 0$ implies that

$$\tilde{Q}_1(t_1', a_1^*(\underline{t}_1)) = \hat{\pi}_1(t_1')[\hat{v}_2(t_2) \geq 1]$$

So equation (3.58) is equivalent to

$$Q_1(t_1', a_1^*(t_1')) > Q_1(t_1', a_1^*(\underline{t}_1)) \tag{3.60}$$

The rest of the proof has the following structure: I start by looking at the incentive compatibility constraint for t_1' to not want to deviate to $a_1^*(\underline{t}_1)$ and show that

combining it with equation (3.60) implies a relationship between the relative probability of receiving the public good and the relative transfers between $a_1^*(t_1')$ and $a_1^*(\underline{t}_1)$. Then using the properties of $\hat{t}_1$ I show that relationship implies equation (3.59). By incentive compatibility of M for t_1', we have

$$U_1(t_1', a_1^*(t_1')) \geq U_1(t_1', a_1^*(\underline{t}_1)) \tag{3.61}$$

which is equivalent to

$$(Q_1(t_1', a_1^*(t_1')) - Q_1(t_1', a_1^*(\underline{t}_1))) \cdot \hat{v}_1(t_1') \geq E_{\hat{\pi}_1(t_1')}\ [\tau_1(a_1^*(t_1'), a_2^*(t_2)) - \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2))]\,. \tag{3.62}$$

Note that by equation (3.60), the left side of this equation is strictly positive. By M as efficient as $\tilde{M}$, both $a_1^*(t_1')$ and $a_1^*(\underline{t}_1)$ must produce the public good when $\hat{v}_2(t_2) \geq 1$. So I can rewrite equation (3.62) as

$$\left(E_{\hat{\pi}_1(t_1')}\ \left[q(a_1^*(t_1'), a_2^*(t_2)) \cdot I_{\hat{v}_2(t_2)<1}\right] + \hat{\pi}_1(t_1')[\hat{v}_2(t_2) \geq 1] - \hat{\pi}_1(t_1')[\hat{v}_2(t_2) \geq 1]\right) \cdot \hat{v}_1(t_1')$$

$$\geq E_{\hat{\pi}_1(t_1')}\ [\tau_1(a_1^*(t_1'), a_2^*(t_2)) - \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2))]$$

which simplifies to

$$E_{\hat{\pi}_1(t_1')}\ \left[q(a_1^*(t_1'), a_2^*(t_2)) \cdot I_{\hat{v}_2(t_2)<1}\right] \cdot \hat{v}_1(t_1') \geq E_{\hat{\pi}_1(t_1')}\ [\tau_1(a_1^*(t_1'), a_2^*(t_2)) - \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2))] \tag{3.63}$$

As with equation (3.62) the left side of equation (3.63) is strictly greater than 0. I now show that equation (3.63) combined with the conditions on $\hat{t}_1$ imply equation (3.59). Condition (A3) implies that

$$E_{\hat{\pi}_1(\hat{t}_1)}\ \left[q(a_1^*(t_1'), a_2^*(t_2)) \cdot I_{\hat{v}_2(t_2)<1}\right] \geq E_{\hat{\pi}_1(t_1')}\ \left[q(a_1^*(t_1'), a_2^*(t_2)) \cdot I_{\hat{v}_2(t_2)<1}\right] > 0 \tag{3.64}$$

(Again the right side greater than zero is implied by (3.60) .) Condition (A1) implies

$$\hat{v}_1(\hat{t}_1) > \hat{v}_1(t_1') \tag{3.65}$$

while (A4) implies that

$$E_{\hat{\pi}_1(\hat{t}_1)}\left[\tau_1(a_1^*(t_1'), a_2^*(t_2)) - \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2))\right] = E_{\hat{\pi}_1(t_1')}\left[\tau_1(a_1^*(t_1'), a_2^*(t_2)) - \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2))\right]. \tag{3.66}$$

Substituting equations (3.64), (3.65), and (3.66), into (3.63) yields

$$E_{\hat{\pi}_1(\hat{t}_1)}\left[q(a_1^*(t_1'), a_2^*(t_2)) \cdot I_{\hat{v}_2(t_2)<1}\right] \cdot \hat{v}_1(\hat{t}_1) > E_{\hat{\pi}_1(\hat{t}_1)}\left[\tau_1(a_1^*(t_1'), a_2^*(t_2)) - \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2))\right] \tag{3.67}$$

By the same logic that shows equations (3.62) and (3.63) are equivalent, equation (3.67) is equivalent to

$$\left(Q_1(\hat{t}_1, a_1^*(t_1')) - Q_1(\hat{t}_1, a_1^*(\underline{t}_1))\right) \cdot \hat{v}_1(\hat{t}_1) > E_{\hat{\pi}_1(\hat{t}_1)}\left[\tau_1(a_1^*(t_1'), a_2^*(t_2)) - \tau_1(a_1^*(\underline{t}_1), a_2^*(t_2))\right]. \tag{3.68}$$

Equation (3.68) is in turn equivalent to

$$U_1(\hat{t}_1, a_1^*(t_1')) > U_1(\hat{t}_1, a_1^*(\underline{t}_1)) \tag{3.69}$$

which, combined with sublemmas 6 and 7 implies equation (3.59). □

This completes the proof of lemma III.10, as sublemma 8 implies that $Q_1(t_1') = \tilde{Q}_1(t_1)$. □

Lemma III.11 completes the proof of Case 2 by showing that types exist that satisfy the conditions posited in lemma III.10.

Lemma III.11. *There exists types $\hat{t}_1$, $\overline{t}_1$ and $\underline{t}_1$ that satisfy the conditions assumed*

in lemma III.10.

Proof. Valuations can be assigned arbitrarily, so the complicated part of the proof is showing that there exist beliefs that satisfy conditions (A2), (A3) and (A4) to construct $\hat{t}_1$, and (B2), (B3) and (B4) to construct $\bar{t}_1$ and $\underline{t}_1$. The construction of beliefs in both cases uses sublemma 10. Sublemma 9 establishes an important result for the proof of sublemma 10, so I prove it first. Before getting to the sublemmas, I introduce some useful terminology for distinguishing between actions played by types of agent 2 with valuations below and (weakly) above 1.

Actions played by types of agent 2 with valuations above and below 1. First I look at the restrictions put on outcomes by the fact that outcomes must be as efficient under M and $\tilde{M}$ for any agent 2.

Let $A_2^* = \{a_2 \in A_2 \mid \exists t_2 \in T_2, \hat{v}_2(t_2) \geq 1, a_2^*(t_2) = a_2\}$

Let $\underline{A}_2 = A_2 \setminus A_2^*$.

If $\tilde{t}_2$'s valuation is at least one, $\tilde{t}_2$ must have $\tilde{a}_2^*(\tilde{t}_2) =$"pay" and must produce the public good under $\tilde{M}$ for any action by agent 1. If $\tilde{t}_2$'s valuation is less than one, $\tilde{t}_2$ must have $\tilde{a}_2^*(\tilde{t}_2) =$"don't pay" and agent 2 does not produce the public good under $\tilde{M}$ if agent 1 plays "DON'T PAY". If M as efficient as $\tilde{M}$, then it must be that agent 2's action does not produce the public good for some action(s) by agent 1 under M (because it would be inefficient when agent 1's valuation is 0), whereas it still must be true that when agent 2's valuation is at least 1 then agent 2 produces the public good for any action by agent 1. Therefore for all $\tilde{t}_2 \in T_2$,

$$\begin{cases} \hat{v}_2(\tilde{t}_2) \geq 1 & \Leftrightarrow \quad a_2^*(\tilde{t}_2) \in A_2^* \\ \hat{v}_2(\tilde{t}_2) < 1 & \Leftrightarrow \quad a_2^*(\tilde{t}_2) \in \underline{A}_2 \end{cases} \tag{3.70}$$

Now I show the following relationship between actions in A_2^* and $\underline{A}_2$. For every action in A_2^*, I can find an action in $\underline{A}_2$ such that I can change an agent 1 type's beliefs by shifting weight from one action onto the other action without changing expected transfers for any action agent 1 might take. This will be crucial in the construction used in sublemma 10.

Sublemma 9. For any action $\hat{a}_2 \in A_2^*$ there exists an action $\underline{a}_2(\hat{a}_2) \in \underline{A}_2$ such that $\forall a_1 \in A_1$,

$$\tau_1(a_1, \hat{a}_2) = \tau_1(a_1, \underline{a}_2(\hat{a}_2)) \tag{3.71}$$

Proof. We can use budget balance to rewrite equation (3.71) in terms of transfers for agent 2. Then it becomes, $\forall a_1 \in A_1$,

$$q(a_1, \hat{a}_2) - \tau_2(a_1, \hat{a}_2) = q(a_1, \underline{a}_2(\hat{a}_2)) - \tau_2(a_1, \underline{a}_2(\hat{a}_2)) \tag{3.72}$$

which is what I will prove.

I show the result in the following steps (the result to be proved is in bold):

1. $\hat{a}_2$ **is a strict best response for some type** $\hat{t}_2$ **with valuation greater than** 1**.** By the assumption that all actions are somewhere strictly preferred in equilibrium, there exists a type $\hat{t}_2$ such that $\hat{t}_2$ strictly prefers $\hat{a}_2$ in equilibrium. $\hat{a}_2 \in A_2^*$ implies that $\hat{v}_2(\hat{t}_2) \geq 1$. In fact it must be that $\hat{v}_2(\hat{t}_2) > 1$, because otherwise types with the same beliefs as $\hat{t}_2$ and valuations less than but arbitrarily close to 1 would strictly prefer to play $\hat{a}_2 \in A_2^*$, which would violate equation (3.70).

2. $\hat{a}_2$ **is a strict best response for an open set of types** t_2^ϵ **with valuation**

greater than 1**.** By the focus lemma, and that $\hat{t}_2$ strictly prefers $\hat{a}_2$, there exists some $\epsilon > 0$ such that for any t_2^ϵ where $\hat{v}_2(t_2^\epsilon) = \hat{v}_2(\hat{t}_2)$ and for all $E \subset T_1$, $|\hat{\pi}_2(t_2^\epsilon)[E] - \hat{\pi}_2(\hat{t}_2)[E]| < \epsilon$, t_2^ϵ also strictly prefers $\hat{a}_2$ in equilibrium. Let T_2^ϵ be the set of such t_2^ϵ.

3. $\hat{a}_2$ **is a (weak) best response for an open set of types** $\tilde{t}_2^\epsilon$ **with valuation** 1**.** Consider any such $t_2^\epsilon \in T_2^\epsilon$. $\hat{a}_2 \in A_2^*$ implies that $\hat{a}_2$ must produce the public good under $\tilde{M}$ for any action by agent 1. Therefore $Q_2(t_2^\epsilon, \hat{a}_2) = 1$. For any valuation $\tilde{v} \geq 1$, let $t_{2,\tilde{v}}^\epsilon$ be the type with beliefs $\hat{\pi}_2(t_{2,\tilde{v}}^\epsilon) = \hat{\pi}_2(t_2^\epsilon)$ and valuation $\hat{v}_2(t_{2,\tilde{v}}^\epsilon) = \tilde{v}$. Let $\tilde{t}_2^\epsilon$ be the type where $\hat{\pi}_2(\tilde{t}_2^\epsilon) = \hat{\pi}_2(t_2^\epsilon)$ and $\hat{v}_2(\tilde{t}_2^\epsilon) = 1$. Then $a_2^*(\tilde{t}_2^\epsilon) \in A_2^*$ which implies $Q_2(\tilde{t}_2^\epsilon, a_2^*(\hat{t}_2^\epsilon)) = 1$.

   Because t_2^ϵ and $\tilde{t}_2^\epsilon$ have the same beliefs, and $Q_2(t_2^\epsilon) = Q_2(\tilde{t}_2^\epsilon) = 1$, it must be that their equilibrium actions are optimal in equilibrium for each other. Therefore $\hat{a}_2$ is a best reply for $\tilde{t}_2^\epsilon$.

   Let $\tilde{T}_2^\epsilon$ be the set of $\tilde{t}_2^\epsilon$ constructed as above. Let $\tilde{t}_2 \in \tilde{T}_2^\epsilon$ be the type such that $\hat{\pi}_2(\tilde{t}_2) = \hat{\pi}_2(\hat{t}_2)$ and $\hat{v}_2(\tilde{t}_2) = 1$.

4. **There exists a** $\tilde{a}_2 \in \underline{A}_2$ **that is also a best response for** $\tilde{t}_2^\epsilon$**.** Take a sequence of valuations $\{v_n\}$ with $v_n < 1$ for all n and such that the limit of the sequence is 1. Let $\tilde{t}_{2,v_n}$ be the type of agent 2 where $\hat{\pi}_2(\tilde{t}_{2,v_n}) = \hat{\pi}_2(\tilde{t}_2)$ and $\hat{v}_2(\tilde{t}_{2,v_n}) = v_n$. Because A_2 is finite, it must be the case that the sequence $a_2^*(\tilde{t}_{2,v_n})$ has a constant subsequence. Call the action in the constant subsequence $\tilde{a}_2$. Then there exist types with the same beliefs as $\tilde{t}_2$ and valuations less than but arbitrarily close to 1 that play $\tilde{a}_2$ in equilibrium. It is then immediate that $\tilde{a}_2$ must be a (weak) best reply for $\tilde{t}_2$.

5. **Then** $\tilde{t}_2$ **indifferent between** $\hat{a}_2$ **and** $\tilde{a}_2$**, and** $\hat{a}_2$ **a best response for all**

$\tilde{t}_2^\epsilon \in \tilde{T}_2^\epsilon$ implies equation (3.72). $\tilde{t}_2$ indifferent between $\hat{a}_2$ and $\tilde{a}_2$ implies

$$E_{\hat{\pi}_2(\hat{t}_2)}\left[q(a_1,\hat{a}_2)-\tau_2(a_1,\hat{a}_2)\right]=E_{\hat{\pi}_2(\hat{t}_2)}\left[q(a_1,\tilde{a}_2)-\tau_2(a_1,\tilde{a}_2)\right] \tag{3.73}$$

Take any $a_1 \in A_1$, and a $\tilde{t}_1$ such that $a_1^*(\tilde{t}_1)=a_1$. For a small enough ϵ we can construct $\tilde{t}_2^\epsilon \in \tilde{T}_2^\epsilon$ such that

$$\hat{\pi}_2(\tilde{t}_2^{\epsilon'})[E]=(1-\epsilon')\cdot\hat{\pi}_2(\tilde{t}_2^\epsilon)[E]+\epsilon'\cdot I_{\tilde{t}_1\in E}. \tag{3.74}$$

Then $\tilde{t}_2^{\epsilon'} \in \tilde{T}_2^\epsilon$ and $\hat{a}_2$ a best reply for all types in $\tilde{T}_2^\epsilon$ implies

$$U_2(\tilde{t}_2^\epsilon,\hat{a}_2)\geq U_2(\tilde{t}_2^\epsilon,\tilde{a}_2) \tag{3.75}$$

which is equivalent to

$$(1-\epsilon)\cdot U_2(\tilde{t}_2^\epsilon,\hat{a}_2)+\epsilon\cdot(q(a_1,\hat{a}_2)-\tau_2(a_1,\hat{a}_2))$$

$$\geq(1-\epsilon)\cdot U_2(\tilde{t}_2^\epsilon,\tilde{a}_2)+\epsilon'\cdot(q(a_1,\tilde{a}_2)-\tau_2(a_1,\tilde{a}_2)) \tag{3.76}$$

which is equivalent to

$$q(a_1,\hat{a}_2)-\tau_2(a_1,\hat{a}_2)\geq q(a_1,\tilde{a}_2)-\tau_2(a_1,\tilde{a}_2). \tag{3.77}$$

Because equation (3.77) must hold for all $a_1 \in A_1$, then equations (3.77) and (3.73) together imply that for all $a_1 \in A_1$,

$$q(a_1,\hat{a}_2)-\tau_2(a_1,\hat{a}_2)=q(a_1,\tilde{a}_2)-\tau_2(a_1,\tilde{a}_2). \tag{3.78}$$

which is the same as equation (3.72), and therefore, as explained above, is equivalent to proving the lemma. □

The next sublemma uses the previous result to show that beliefs can be constructed that have the properties we need to find suitable beliefs for $\hat{t}_1$, $\bar{t}_1$, and $\underline{t}_1$.

Sublemma 10. For any $\pi_1' \in \Delta(T_2)$, if $\pi_1'[\hat{v}_2(\tilde{t}_2) < 1] \equiv \alpha < 1$ then for any $\delta \in (\alpha, 1]$ there exists a π_1^δ such that

$$\pi_1^\delta[\hat{v}_2(t_2) < 1] = \delta \tag{3.79}$$

for all $a_1 \in A_1$,

$$E_{\pi_1^\delta}\tau_1(a_1, a_2^*(t_2)) = E_{\pi_1'}\tau_1(a_1, a_2^*(t_2)) \tag{3.80}$$

and $\forall a_2 \in A_2$

$$\pi_1^\delta[a_2^*(\tilde{t}_2) = a_2,\ \hat{v}_2(\tilde{t}_2) < 1] \geq \pi_1'[a_2^*(\tilde{t}_2) = a_2,\ \hat{v}_2(\tilde{t}_2) < 1]. \tag{3.81}$$

Proof. First I prove equation (3.80). I start by constructing some beliefs π_1^1 that I will use to define π_1^δ. Relabel $A_2 = \{a_2^1, ..., a_2^{|A_2|}\}$. For $1 \leq n \leq |A_2|$ find t_2^n such that $a_2(t_2^n) = a_2^n$.

For $a_2^n \in A_2^*$, define $\underline{a}_2(a_2^n)$ as in the lemma 9. For $a_2^n \in \underline{A}_2$, define $\underline{a}_2(a_2^n) \equiv a_2^n$. This construction ensures that for all $a_2 \in A_2$,

$$\underline{a}_2(a_2^n) \in \underline{A}_2. \tag{3.82}$$

Define the following beliefs for agent 1:

$$\pi_1^1(t_2^n) = \sum_{m=1}^{|A_2|} I_{\underline{a}_2(a_2^m)=a_2^n} \cdot \pi_1'[a_2^*(\tilde{t}_2) = a_2^m]. \tag{3.83}$$

Note that equation (3.83) implies

$$\pi_1^1[\hat{v}_2(\tilde{t}_2) < 1] = 1. \tag{3.84}$$

Now I want to show that for any $a_1 \in A_1$, the expected transfer given beliefs $\hat{\pi}_1^1$ and π_1' are equal. The following calculation shows that this is true:

$$E_{\pi_1^1}\tau_1(a_1, a_2^*(t_2)) = \sum_{m=1}^{|A_2|} \tau_1(a_1, a_2^n) \cdot \pi_1^1(t_2^n) \tag{3.85}$$

$$= \sum_{m=1}^{|A_2|} \tau_1(a_1, a_2^n) \left(\sum_{m=1}^{|A_2|} I_{\underline{a}_2(a_2^m)=a_2^n} \cdot \pi_1'[a_2^*(\tilde{t}_2) = a_2^m] \right) \tag{3.86}$$

$$= \sum_{n=1}^{|A_2|} \sum_{m=1}^{|A_2|} \tau_1(a_1, a_2^n) I_{\underline{a}_2(a_2^m)=a_2^n} \cdot \pi_1'[a_2^*(\tilde{t}_2) = a_2^m] \tag{3.87}$$

$$= \sum_{m=1}^{|A_2|} \pi_1'[a_2^*(\tilde{t}_2) = a_2^m] \sum_{n=1}^{|A_2|} \tau_1(a_1, a_2^n) I_{\underline{a}_2(a_2^m)=a_2^n} \tag{3.88}$$

which by sublemma 9 is equivalent to

$$= \sum_{m=1}^{|A_2|} \pi_1'[a_2^*(\tilde{t}_2) = a_2^m] \cdot \tau_1(a_1, a_2^m) \tag{3.89}$$

$$= E_{\pi_1'}\tau_1(a_1, a_2^*(t_2)). \tag{3.90}$$

Now I will construct π_1^δ. Let π_1^δ be defined by, for all $E \subset T_2$,

$$\pi_1^\delta(E) = \frac{1-\delta}{1-\alpha}\pi_1'(E) + \frac{\delta-\alpha}{1-\alpha}\pi_1^1(E) \tag{3.91}$$

Recall that $\alpha \equiv \pi_1'[\hat{v}_2(t_2) < 1]$, and equation (3.84) shows $\pi_1^1[\hat{v}_2(t_2) < 1] = 1$. Therefore,

$$\pi_1^\delta[\hat{v}_2(t_2) < 1] = \frac{1-\delta}{1-\alpha}\alpha + \frac{\delta-\alpha}{1-\alpha} \cdot 1 = \delta \tag{3.92}$$

which proves equation (3.79), and

$$E_{\pi_1^\delta}\tau_1(a_1, a_2^*(t_2)) \tag{3.93}$$

$$= \frac{1-\delta}{1-\alpha} \cdot E_{\pi_1'}\tau_1(a_1, a_2(t_2)) + \frac{\delta-\alpha}{1-\alpha}E_{\pi_1^1}\tau_1(a_1, a_2^*(t_2)) \tag{3.94}$$

$$= E_{\pi_1'}\tau_1(a_1, a_2^*(t_2)) \tag{3.95}$$

which proves equation (3.80).

To prove equation (3.81), observe that for all $a_2^n \in \underline{A}_2$, $\underline{a}_2(a_2^n) = a_2^n$, which implies:

$$\pi_1^1[a_2^*(\tilde{t}_2) = a_2^n] = \sum_{m=1}^{|A_2|} \left(I_{\underline{a}_2(a_2^m)=a_2^n} \cdot \pi_1'[a_2^*(\tilde{t}_2) = a_2^m] \right) \geq \pi_2'[a_2^*(\tilde{t}_2) = a_2^n]. \tag{3.96}$$

Furthermore, by equation (3.70) it must be the case that $a_2^*(\tilde{t}_2) = a^n \in \underline{A}_2$ implies $\hat{v}_2(\tilde{t}_2) < 1$. Therefore,

$$\pi_1^1[a_2^*(\tilde{t}_2) = a_2^n, \hat{v}_2(\tilde{t}_2) < 1] = \pi_1^1[a_2^*(\tilde{t}_2) = a_2^n] \tag{3.97}$$

and

$$\pi_1'[a_2^*(\tilde{t}_2) = a_2^n, \hat{v}_2(\tilde{t}_2) < 1] = \pi_1'[a_2^*(\tilde{t}_2) = a_2^n] \tag{3.98}$$

and equations (3.96), (3.97) and (3.98) imply that for all $a_2^n \in \underline{A}_2$,

$$\pi_1^1[a_2^*(\tilde{t}_2) = a_2^n, \hat{v}_2(\tilde{t}_2) < 1] \geq \pi_1'[a_2^*(\tilde{t}_2) = a_2^n, \hat{v}_2(\tilde{t}_2) < 1]. \tag{3.99}$$

Furthermore, equation (3.70) implies that for all $a_2^n \in A_2^*$,

$$\pi_1^1[a_2^*(\tilde{t}_2) = a_2^n, \hat{v}_2(\tilde{t}_2) < 1] = \pi_1'[a_2^*(\tilde{t}_2) = a_2^n, \hat{v}_2(\tilde{t}_2) < 1] = 0. \tag{3.100}$$

Together equations (3.99) and (3.100) imply that for all $a_2 \in A_2$,

$$\pi_1^1[a_2^*(\tilde{t}_2) = a_2, \hat{v}_2(\tilde{t}_2) < 1] \geq \pi_1'[a_2^*(\tilde{t}_2) = a_2, \hat{v}_2(\tilde{t}_2) < 1]. \tag{3.101}$$

and because π_1^δ is a convex combination of $\hat{\pi}_1^1$ and $\hat{\pi}_1'$ it must be that

$$\pi_1^\delta[a_2^*(\tilde{t}_2) = a_2, \hat{v}_2(\tilde{t}_2) < 1] \geq \pi_1'[a_2^*(\tilde{t}_2) = a_2, \hat{v}_2(\tilde{t}_2) < 1] \tag{3.102}$$

which proves equation (3.81). □

Sublemma 11. Let $\pi_1' = \hat{\pi}_1(t_1')$. Then using the construction in sublemma 10, the following types

- $\hat{t}_1$ with a valuation $\hat{v}_1(\hat{t}_1) > \max\{\hat{v}_1(t_1^*), 1\}$ and $\hat{\pi}_1(\hat{t}_1) = \pi_1^\delta$ where $\delta = \frac{1}{\hat{v}_1(\hat{t}_1)}$
- $\bar{t}_1$ with valuation $\hat{v}_1(\bar{t}_1) = \hat{v}_1(\hat{t}_1)$ and $\hat{\pi}_1(\bar{t}_1) = \pi_1^1$
- $\underline{t}_1$ with valuation $\hat{v}_1(\underline{t}_1) = 0$ and $\hat{\pi}_1(\underline{t}_1) = \hat{\pi}_1(\bar{t}_1)$

satisfy conditions (A1)-(A4), (B1)-B(4), and (C1)-(C2) respectively.

Proof. By construction $\hat{t}_1$ satisfies (A1) and (A2), and sublemma 10 implies $\hat{t}_1$ satisfies (A3) and (A4). Therefore a $\hat{t}_1$ that satisfies (A1)-(A4) exists.

Similarly, by construction $\bar{t}_1$ satisfies (B1) and (B2), and sublemma 10 implies that $\bar{t}_1$ satisfies (B3) and (B4).

$\underline{t}_1$ satisfies (C1) and (C2) by construction.

□

Sublemma 11 completes the proof of lemma III.11 by providing explicit types that satisfy the conditions posited in lemma III.10. □

Together lemmas III.10 and III.11 complete the proof of case 2 by showing that $Q_1(t_1') = \tilde{Q}_1(t_1')$. □

Together, the proofs of Case 1 and Case 2 complete the proof of proposition III.7, which is equivalent to proposition III.5. Therefore no M and a^* satisfying the conditions in proposition III.5 improves on $\tilde{M}$ and $\tilde{a}^*$. □

3.6 Conclusion

In chapter 2 I introduced the concept of improvability as a criterion for assessing whether a public good mechanism can be considered efficient. While I showed the existence of unimprovable mechanisms in that paper, I presented no examples. This paper presents an example of an unimprovable mechanism, but also highlights some of the difficulties in applying the unimprovability concept; although the mechanism is simple the proof is quite involved, and some extra restrictions on the class of mechanism are required. Methods to characterize the class of unimprovable mechanisms, and to determine whether a given mechanism is improvable or unimprovable, are areas for further research.

CHAPTER IV

Robust Mechanism Design and Dominant Strategy Voting Rules

4.1 Introduction

Economic outcomes depend not only on market processes but also on political processes. Economists have therefore a long-standing interest in political decision making. Political decisions are often made through voting procedures. Consequently it is interesting to investigate which voting procedures perform *well* in the sense of helping to achieve some measure of economic welfare. One methodology that can be used to address this question is the theory of *mechanism design.* In this paper we consider the design of voting rules from the perspective of the theory of mechanism design.

Our starting point is a classic result on voting rules, due to Alan Gibbard (1973) and Mark Satterthwaite (1975). According to this result the only dominant strategy voting rules for three or more alternatives are dictatorial voting rules. Gibbard and Satterthwaite assumed the number of alternatives to be finite. Preferences were modeled as complete and transitive orders of the set of alternatives. For every voter the range of relevant preferences was taken to be the set of *all* possible preferences over the alternatives (the *full domain assumption*). Gibbard and Satterthwaite then asked

whether it is possible to construct a game form[1] that determines which alternative is chosen as a function of the strategies chosen by the voters, such that each voter has a dominant strategy whatever this voter's preferences are. A dominant strategy was defined to be a strategy that is always a best reply to each of the other voters' strategy combinations. Gibbard and Satterthwaite showed that the only game forms that offer each voter for all preferences a dominant strategy are game forms that leave the choice of the outcome to just one individual, the dictator.[2]

The motivation for considering dominant strategy game forms is not always articulated in the literature. However, one explanation for the appeal that dominant strategy mechanisms have for researchers is that dominant strategies predict rational voters' behavior without relying on any assumption about the voters' beliefs about each others' preferences or behavior. If a voter does not have a dominant strategy, then that voter's optimal choice depends on his beliefs about other voters' behavior which in turn may be derived from beliefs about other voters' preferences. It seems attractive to bypass such beliefs, and to construct a game form in which a prediction can be made that is independent of beliefs.

On closer inspection, this argument can be seen to consist of two parts:

> (A) *The design of a good game form for voting should not be based on specific assumptions about voters' beliefs about each other.*
>
> (B) *A good game form for voting should allow us to predict rational voters' choices uniquely from their preferences, without making specific assumptions about these voters' beliefs about each other.*[3]

These two parts are logically independent. Part (A) seems more convincing: often

[1] We use the terms *game form* and *mechanism* synonymously.

[2] The literature that builds on Gibbard and Satterthwaite's seminal work is voluminous. For a recent survey see Barberà (2010).

[3] Blin and Satterthwaite (1977) emphasize the interpretation of the Gibbard Satterthwaite theorem as a result about voting procedures in which each voter's choice depends only on their preferences, and not on their beliefs about others' preferences.

voting schemes are constructed long before the precise context in which they will be used is known. It seems wise not to make any special assumptions about agents' knowledge about each other. Part (B) can perhaps be motivated by the idea that game forms in which voters' behavior can be uniquely predicted independent of their beliefs are simpler than game forms in which each voter's optimal choice depends on the voter's beliefs about other voters, but this point seems less compelling. The implicit idea of simplicity is just one of several conceivable notions of simplicity.

In this paper we present an investigation of the theory of voting rules that is based on the first part of the two part argument described above, but not on the second part. In other words, we examine game forms for voting without making assumptions about voters' beliefs about each other, but we do not restrict attention to game forms for which voters' equilibrium strategies are independent of voters' beliefs. Using the terminology of game theory, the fact that we do not make any assumptions about voters' beliefs about each other is reflected by the fact that we analyze any proposed game form for *all* possible type spaces. For each type space we look for a Bayesian equilibrium of the given game form for that type space.[4] However, we do not require each voter's choice, for a given preference of that voter, to be the same for all type spaces.

One of the two main findings of this paper is that a mechanism designer who evaluates voting rules using the Pareto criterion can improve on dictatorial mechanisms even when not making any assumption about voters' beliefs about each other. A Pareto improvement on dictatorship is possible in our framework when we consider *random* dictatorship where all agents have a positive probability of being dictator. To explain our results more fully, we need to briefly describe the set-up of our paper.

In order to be able to use the notion of *Bayesian equilibrium* we use a framework

[4]For the definitions of *type space* and *Bayesian equilibrium* see Fudenberg and Tirole (1991, pp. 213-215).

that is slightly different from the framework that Gibbard and Satterthwaite used. We model voters' attitudes towards risk, adopting the assumption that voters evaluate risky prospects according to von Neumann Morgenstern utility theory. It then seems natural to allow voting rules to map profiles of von Neumann Morgenstern utility functions into probability distributions over outcomes. The first question that arises is whether a version of Gibbard and Satterthwaite's theorem holds for the setting just described. This question has been answered affirmatively by Aanund Hylland in 1980 in the unpublished (Hylland, 1980). When voters have von Neumann Morgenstern utilities, and lotteries are allowed as outcomes, then the only game forms that offer each agent always a dominant strategy, and that pick an alternative if it is unanimously preferred by all agents, are random dictatorships.[5] In random dictatorships each voter gets to be dictator with a probability p_i that is independent of all preferences. If voter i is dictator, then the outcome that voter i ranks highest is chosen.

We can now state the two main results of this paper. Both results address whether there are game forms such that for all finite type spaces, there is at least one Bayesian equilibrium of the game form that yields all voters' types the same expected utility, and in some type spaces, for some voters' types, strictly higher expected utility than random dictatorship. Obviously, the answer to this question can be positive only when each voter's probability of being dictator is strictly less than one. In our first main result we show that in this case the answer to our question is indeed positive,[6] provided that we consider *interim* expected utility, that is, each voter's expected

[5]This result is Theorem 1* in Hylland (1980). It is also Theorem 1 in Dutta et. al. (2007) (see also Dutta et. al., 2008) where an alternative proof is provided. Another proof is in Nandeibam (2004).

[6]The game form that we use to prove our first main result is almost identical to the *Full Consensus or Random Ballot Fall-Back* game form that Heitzig and Simmons (2010) have introduced. While their motivation, like ours, is to consider voting systems that are more flexible than dictatorial voting systems, and that allow for compromises, the focus of their formal analysis is on complete information, correlated equilibria that are in some sense coalition proof. In this paper the focus is on analyzing Bayesian equilibria in arbitrary, finite type spaces.

utility is calculated when that voter's type is known, but the other voters' types are not yet known.[7] If an *ex post* perspective is adopted instead, that is, if voters' expected utility is considered conditional on the vector of *all* voters' types, then no voting game form Pareto improves on random dictatorship. This is our second main result. Our first main result thus indicates that a robust analysis of voting schemes can lead to more positive results if the requirement that voters' optimal strategies are independent of their beliefs is abandoned. Our second main result shows that such positive results depend on the details of how each voter's expected utility is evaluated.

Our approach is related to Bergemann and Morris' (2005) work on robust mechanism design. They consider, as we do, Bayesian equilibria of mechanisms on *all* type spaces. Bergemann and Morris seek conditions under which the Bayesian implementability of a social choice correspondence on all type spaces implies dominant strategy implementability (or, more generally, implementability in *ex post equilibria*). The conditions that they find apply to *separable environments* the prime example of which are environments in which each agent's utility depends on some physical allocation and this agent's monetary transfer. Bergemann and Morris point out (2005, Section 6.3) that in non-separable environments, such as environments without transferrable payoffs considered by Gibbard and Satterthwaite, dominant strategy implementability may be a stronger requirement than Bayesian implementability on all type spaces.[8] Bergemann and Morris do not consider the problem of comparing different mechanisms from an efficiency or welfare point of view. Such comparisons are a focus in our work.

The approach of this paper are also closely related to chapter 2 of this dissertation, which analyzes the problem of designing a mechanism for public goods. Like we do in this chapter, chapter 2 considers the performance of different mechanisms on all

[7] The notions of interim and ex post efficiency are due to Holmström and Myerson (1983).

[8] The discussion paper version ((Bergemann and Morris, 2003) of Bergemann and Morris (2005) includes a general characterization of Bayesian implementability on all type spaces, however we do not make use of this characterization.

type spaces. That chapter focuses on an ex post perspective, and demonstrates that a mechanism designer can improve efficiency using a more flexible mechanism than a dominant strategy mechanism. In this chapter, by contrast, when considering the ex post perspective, we find that no mechanism can improve on dominant strategy mechanisms.

The spirit of our work in this paper is also related to Börgers (1991) who showed, in the Gibbard-Satterthwaite framework, the existence of mechanisms for which the outcomes that result if all players chose a strategy from their sets of *undominated strategies* are Pareto efficient, and (in a sense defined in that paper) less biased than the outcomes of dictatorship. The set of undominated strategies is equal to the set of expected utility maximizing strategies that a rational agent might choose if one considers all possible beliefs. Thus, implicitly, Börgers (1991) considered implementation on all type spaces with belief-dependent strategies, and contrasted this with Gibbard and Satterthwaite's dominant strategy requirement. However, Börgers used a framework in which agents' preferences were modeled using ordinal preferences rather than von Neumann Morgenstern utilities. Moreover, his approach can be considered an *implementation* approach, as he considered *all* undominated strategies, whereas our approach here is a *mechanism design* approach in the sense that we study for every type space *some* equilibrium, but not *all* Bayesian equilibria. We leave the further exploration of the implementation approach in our framework to future research.

Section 4.2 explains the model and the definitions used in this paper. In Section 4.3 we adapt Hylland's theorem on random dictatorship to our setting. In Section 4.4 we explain how we relax the requirement that voters' choices, for given preferences, are the same in all type spaces. Sections 4.5 and 4.6 contain our two main results. Section 4.7 concludes.

4.2 The Voting Problem

There are n agents: $i \in I = \{1, 2, \ldots, n\}$. The agents have to choose one alternative from a finite set A of alternatives. We assume that A has at least three elements. The set of all probability distributions over A is $\Delta(A)$, where for $\delta \in \Delta(A)$ we denote by $\delta(a) \in [0, 1]$ the probability that δ assigns to alternative a. The agents are commonly known to be expected utility maximizers. We denote agent i's von Neumann Morgenstern utility function by $u_i : A \to \mathbb{R}$. We assume that $a \neq b \Rightarrow u_i(a) \neq u_i(b)$, i.e., there are no indifferences. We define the expected utility for probability distributions $\delta \in \Delta(A)$ by $u_i(\delta) = \sum_{a \in A} u_i(a)\delta(a)$.

A mechanism designer has a, possibly incomplete, ranking of the alternatives in A that may depend on the agents' utility functions. We shall be more specific about the designer's objectives later. The mechanism designer does not know the agents' utility functions, nor does she know what the agents believe about each other. To implement an outcome that potentially depends on the agents' utility functions the mechanism designer asks the agents to play a *game form.*

Definition IV.1. A *game form* $G = (S, x)$ consists of:

(i) a set $S \equiv \prod_{i \in I} S_i$ where for every $i \in I$ the set S_i is non-empty and finite;

(ii) a function $x : S \to \Delta(A)$.

The set S_i is the set of (pure) strategies available to agent i in the game form G. We focus on finite sets of pure strategies, while allowing mixed strategies, to ease exposition. Our results also hold when the sets S_i of pure strategies are allowed to be infinite. The function x assigns to every combination of pure strategies s the, potentially stochastic, outcome $x(s)$ that is implemented when agents choose that combination of pure strategies. We write $x(s, a)$ for the probability that $x(s)$ assigns to alternative a.

Once the mechanism designer has announced a game form, the agents choose simultaneously and independently their strategies. Because the agents don't necessarily know each others' utility functions or beliefs, this game may be a game of incomplete information. A hypothesis about the agents' utility functions and their beliefs about each other can be described by specifying a *type space.*

Definition IV.2. A *type space* $\mathcal{T} = (T, \pi, u)$ consists of:

(i) a set $T \equiv \prod_{i \in I} T_i$, where for every $i \in I$ the set T_i is non-empty and finite;

(ii) an array $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ of functions $\pi_i : T_i \to \Delta(T_{-i})$ where $\Delta(T_{-i})$ is the set of all probability distributions over $T_{-i} \equiv \prod_{j \neq i} T_j$;

(iii) an array $u = (u_1, u_2, \ldots, u_n)$ of functions $u_i : T_i \times A \to \mathbb{R}$ such that $a \neq b \Rightarrow u_i(t_i, a) \neq u_i(t_i, b)$ for all $t_i \in T_i$.

The set T_i is the set of types of agent i. Agent i privately observes his type. The function π_i describes for every type $t_i \in T_i$ the beliefs that agent i has about the other agents' types when agent i himself is of type t_i. We write $\pi_i(t_i, t_{-i})$ for the probability that type t_i assigns to the other players types being t_{-i}. The function $u_i(t_i)$ describes player i's utility when i is of type t_i. We write $u_i(t_i, a)$ for the utility that $u_i(t_i)$ assigns to alternative a. The utility functions $u_i(t_i)$ satisfy the assumption that we introduced earlier that there are no indifferences.[9]

In Definition V.2 beliefs are subjective. There may or may not be a common prior for a particular type space. Different agents' beliefs may be incompatible with each other in the sense that one agent may attach probability one to an event to which another agent attaches probability zero. Observe also that we assume type spaces to be finite. We thus avoid technical difficulties associated with infinite type spaces.

[9]Observe that we suppress in the notation the dependence of π_i and u_i on the type space $\mathcal{T}$. We are not aware of any confusion that might arise from this simplification of our notation.

We assume that the mechanism designer has no knowledge of the agents' utility functions or their beliefs. Therefore, the mechanism designer regards all type spaces as possible descriptions of the environment in which the agents find themselves. We denote the set of all type spaces by Υ.

The mechanism designer proposes to agents how they might play the game. She may propose to the agents to randomize. For $i \in I$ we denote by $\Delta(S_i)$ the set of all probability distributions on S_i. For the agents to accept the mechanism designer's proposal, she must propose a *Bayesian equilibrium*. Because the mechanism designer does not know the true type space, she has to propose a *Bayesian equilibrium for every type space*.

Definition IV.3. A *Bayesian equilibrium of game form G for every type space* is an array $\sigma^* = (\sigma_1^*, \sigma_2^*, \ldots, \sigma_n^*)$ such that for every $i \in I$:

(i) σ_i^* is a family of functions $(\sigma_i^*(\mathcal{T}))_{\mathcal{T}\in\Upsilon}$ where for every $\mathcal{T} \in \Upsilon$ the function $\sigma_i^*(\mathcal{T})$ maps the type space T_i corresponding to $\mathcal{T}$ into $\Delta(S_i)$;

and, writing $\sigma_i^*(\mathcal{T}, t_i)$ for the mixed strategy assigned to t_i, and writing $\sigma_i^*(\mathcal{T}, t_i, s_i)$ for the probability that this mixed strategy assigns to $s_i \in S_i$, we have for every $\mathcal{T} \in \Upsilon$, $i \in I$, and $t_i \in T_i$ (where T_i corresponds to $\mathcal{T}$):

(ii) $\sigma_i^*(\mathcal{T}, t_i)$ maximizes the expected utility of type t_i among all mixed strategies in $\Delta(S_i)$, where expected utility for any mixed strategy $\sigma_i \in \Delta(S_i)$ is:

$$\sum_{t_{-i}\in T_{-i}} \pi_i(t_i, t_{-i}) \sum_{s\in S} u_i(t_i, x(s)) \cdot \sigma_i(s_i) \cdot \prod_{j\neq i} \sigma_j^*(\mathcal{T}, t_j, s_j). \tag{4.1}$$

We postulate in this paper a mechanism designer who seeks to further the utility of the agents rather than her own utility. We shall formalize this by assuming that the mechanism designer evaluates different mechanisms and their equilibria using the Pareto criterion. When evaluating the agents' utility for a realized type combination

t the mechanism designer can either only consider the outcomes that result from the mixed strategies prescribed for these types, or she may consider the expected utilities of these types, based on the types' own subjective beliefs. In other words, the mechanism designer may adopt an *ex post* or an *interim* perspective when evaluating agents' utilities. The interim perspective respects agents' own perception of their environment. The ex post perspective has a paternalistic flavor. On the other hand, for example when agents' beliefs are incompatible with each other, the mechanism designer may be justified in discarding agents' beliefs, on the basis that at least some of them have to be wrong, as agents themselves will discover at some point. Thus neither the interim nor the ex post perspective are clearly preferable. We pursue both perspectives in this paper.

Definition IV.4. The game form G and the Bayesian equilibrium for all type spaces σ^* *interim Pareto dominate* the game form $\tilde{G}$ and the Bayesian equilibrium for all type spaces $\tilde{\sigma}^*$ if for all $\mathcal{T} \in \Upsilon$, $i \in I$, and $t_i \in T_i$:

$$\begin{aligned} &\sum_{t_{-i} \in T_{-i}} \pi_i(t_i, t_{-i}) \sum_{s \in S} u_i(t_i, x(s)) \cdot \prod_{j \in I} \sigma_j^*(\mathcal{T}, t_j, s_j) \geq \\ &\sum_{t_{-i} \in T_{-i}} \pi_i(t_i, t_{-i}) \sum_{s \in S} u_i(t_i, \tilde{x}(s)) \cdot \prod_{j \in I} \tilde{\sigma}_j^*(\mathcal{T}, t_j, s_j) \end{aligned} \tag{4.2}$$

with strict inequality for at least one $\mathcal{T} \in \Upsilon$, $i \in I$, and $t_i \in T_i$.

Definition IV.5. The game form G and the Bayesian equilibrium for all type spaces σ^* *ex post Pareto dominate* the game form $\tilde{G}$ and theBayesian equilibrium for all type spaces $\tilde{\sigma}^*$ if for all $\mathcal{T} \in \Upsilon$, $i \in I$, and $t \in T$:

$$\begin{aligned} &\sum_{s \in S} u_i(t_i, x(s)) \cdot \prod_{j \in I} \sigma_j^*(\mathcal{T}, t_j, s_j) \geq \\ &\sum_{s \in S} u_i(t_i, \tilde{x}(s)) \cdot \prod_{j \in I} \tilde{\sigma}_j^*(\mathcal{T}, t_j, s_j) \end{aligned} \tag{4.3}$$

with strict inequality for at least one $\mathcal{T} \in \Upsilon$, $i \in I$, and $t \in T$.

Our main interest in this paper is in exploring how the extent to which the mechanism designer can achieve her objectives depends on the requirements that the Bayesian equilibrium that the mechanism designer proposes has to satisfy. In the next section, we consider a very restrictive requirement. In subsequent sections, we relax this requirement.

4.3 Belief Independent Equilibria: Hylland's Theorem

We begin by exploring the consequences of a restrictive requirement for the Bayesian equilibria that the mechanism designer proposes. This requirement is implicit in the work on dominant strategy mechanism design. It is that equilibria be *belief independent*. Using the notion of belief independent equilibria, we can restate Hylland's version of the Gibbard Satterwhaite theorem in our setting.

Definition IV.6. A game form G and a Bayesian equilibrium of G for every type space, σ^*, are *belief independent* if for all $i \in I, \mathcal{T}, \widetilde{\mathcal{T}} \in \Upsilon, t_i \in T_i$ and $\tilde{t}_i \in \tilde{T}_i$ such that $u_i(t_i) = \tilde{u}_i(\tilde{t}_i)$ we have:

$$\sigma_i^*(\mathcal{T}, t_i) = \sigma_i^*(\widetilde{\mathcal{T}}, \tilde{t}_i), \tag{4.4}$$

where T_i, u_i correspond to $\mathcal{T}$ and $\tilde{T}_i, \tilde{u}_i$ correspond to $\widetilde{\mathcal{T}}$.

The reformulation of Hylland's theorem presented below says that all game forms and belief independent equilibria of these game forms that satisfy two unanimity requirements are random dictatorships. To define the two unanimity requirements and random dictatorship we need some notation. If u is a utility function, we denote by $b(u)$ the element of A that maximizes u, and by $w(u)$ the element of A that

minimizes u.[10]

Definition IV.7. A game form G and a Bayesian equilibrium of G for every type space, σ^*, satisfy

(i) *positive unanimity* if for every $\mathcal{T} \in \Upsilon$, $t \in T$, and $a \in A$ such that $b(u_i(t_i)) = a$ for all $i \in I$, we have:

$$\sum_{s \in S} \prod_{i \in I} \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot x(s, a) = 1; \tag{4.5}$$

(ii) *negative unanimity* if for every $\mathcal{T} \in \Upsilon$, $t \in T$, and $a \in A$ such that $w(u_i(t_i)) = a$ for all $i \in I$, we have:

$$\sum_{s \in S} \prod_{i \in I} \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot x(s, a) = 0. \tag{4.6}$$

Positive and negative unanimity are implied by, but weaker than ex post Pareto efficiency. Next, we provide the formal definition of random dictatorship that we need for our reformulation of Hylland's theorem.

Definition IV.8. A game form G and a Bayesian equilibrium of G for every type space, σ^*, are a *random dictatorship* if there is some $p \in [0, 1]^n$ such that for every $\mathcal{T} \in \Upsilon$, $t \in T$, and $a \in A$:

$$\sum_{s \in S} \prod_{i \in I} \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot x(s, a) = \sum_{\{i \in I : b(u_i(t_i)) = a\}} p_i \tag{4.7}$$

The following is implied by Hylland's theorem.[11]

[10] Recall that we have assumed that there are no indifferences. Therefore, there is a unique element of A that maximizes u, and a unique element of A that minimizes u.

[11] Theorem 1* in Hylland (1980). We use here the version of Hylland's theorem that is Theorem 1 in Dutta et. al. (2007) with the correction in Dutta et. al. (2008).

Proposition IV.9. *A game form G and a Bayesian equilibrium of G for every type space, σ^*, are belief-independent and satisfy positive and negative unanimity if and only if they are a random dictatorship.*

Proof. The "if-part" is obvious. To prove the "only if-part" we derive from G and σ^* a "cardinal decision scheme" in the sense of Definition 1 in Dutta et. al. (2007), and show that this cardinal decision scheme has the properties listed in Theorem 1 in Dutta et. al. (2007) and the correction in Dutta et. al. (2008). It then follows from Theorem 1 in Dutta et. al. (2007) that the cardinal decision scheme is a random dictatorship. This then implies the "only if-part" of our Proposition IV.9.

Denote by $\mathcal{U}$ the set of all utility functions that have the property of no indifferences (see Definition V.2). A cardinal decision scheme is a mapping $\phi : \mathcal{U}^n \to \Delta(A)$. We can derive from G and σ^* a cardinal decision scheme by setting for any $(u_1, u_2, \ldots, u_n) \in \mathcal{U}^n$ and $a \in A$ the probability $\phi(u_1, u_2, \ldots, u_n, a)$ that $\phi(u_1, u_2, \ldots, u_n)$ assigns to a as:

$$\phi(u_1, u_2, \ldots, u_n, a) = \sum_{s \in S} \prod_{i \in I} \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot x(s, a), \tag{4.8}$$

where we can pick any $\mathcal{T} \in \Upsilon$ and any $t \in T$ such that $u_i(t_i) = u_i$ for all $i \in I$. By belief-independence it does not matter which such $\mathcal{T}$ and $t \in T$ we choose. Then ϕ is a cardinal decision scheme as defined in Definition 1 of Dutta et. al. (2007).

We can complete the proof by showing that ϕ has the two properties listed in Theorem 1 of Dutta et. al. (2007) and the additional property listed in the correction Dutta et. al. (2008). The first property is unanimity: If $b(u_i) = a$ for all $i \in I$, then $\phi(u_1, u_2, \ldots, u_n, a) = 1$. This is implied by the assumption that G and σ^* satisfy positive unanimity.

The second property is strategy proofness: If $(u_1, u_2, \ldots, u_n) \in \mathcal{U}^n$ and $u_i' \in \mathcal{U}$, then $u_i(\phi(u_i, u_{-i})) \geq u_i(\phi(u_i', u_{-i}))$, where u_{-i} is the array $(u_1, u_2, \ldots, u_n)$ leaving

out u_i. To prove this we pick $\mathcal{T} \in \Upsilon$, $t_i, t_i' \in T_i$ and $t_{-i} \in \prod_{j \neq i}$ such that $u_i(t_i) = u_i, u_i(t_i') = u_i'$, and $u_j(t_j) = u_j$ for all $j \neq i$. Moreover, $\pi_i(t_i)$ and $\pi_i(t_i')$ place probability 1 on t_{-i}. Then the fact that σ^* is a Bayesian equilibrium of G for the type space $\mathcal{T}$ implies:

$$\sum_{s \in S} u_i(t_i, x(s)) \cdot \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot \prod_{j \neq i} \sigma_j^*(\mathcal{T}, t_j, s_j) \geq$$
$$\sum_{s \in S} u_i(t_i, x(s)) \cdot \sigma_i^*(\mathcal{T}, t_i', s_i) \cdot \prod_{j \neq i} \sigma_j^*(\mathcal{T}, t_j, s_j) \tag{4.9}$$

By the definition of ϕ, this is equivalent to: $u_i(\phi(u_i, u_{-i}) \geq u_i(\phi(u_i', u_{-i}))$, that is, strategy proofness.

The third property, introduced in the correctionDutta et. al. (2008), is a property labelled (*) in Dutta et. al. (2007): If $w(u_i) = a$ for all $i \in I$, then $\phi(u_1, u_2, \ldots, u_n, a) = 0$. This is implied by the assumption that G and σ^* satisfy negative unanimity. □

From now on, when we refer to random dictatorship, we shall mean a specific game form G, and a specific equilibrium σ^* of G for every type space.

Definition IV.10. For any vector $p \in [0,1]^n$ such that $\sum_{i \in I} p_i = 1$ the following game form G and equilibrium σ^* of G for every type space will be referred to as *p-random dictatorship*:

(i) $S_i = A$ for all $i \in I$;

(ii) $x(s, a) = \sum_{\{i \in I : b(u_i(t_i)) = a\}} p_i$ for all $s \in S$ and $a \in A$;

(iii) $\sigma_i^*(\mathcal{T}, t_i, b(u_i(t_i))) = 1$ for all $i \in I$, $\mathcal{T} \in \Upsilon$, and $t_i \in T_i$.

It is immediate that σ^* is a Bayesian equilibrium of G for every type space, and that G and this equilibrium are a random dictatorship. There are other game forms and equilibria that are random dictatorships, but it is without loss of generality to only consider the one described in Definition V.10.

4.4 Consistent Equilibria

Our main interest in this paper is in considering the implications of relaxing the requirement of belief independence for the Bayesian equilibria that the mechanism designer chooses. We do not, however, completely dispense with any link between players' strategies in different type spaces. The Bayesian equilibria that we shall investigate need to satisfy a *consistency* requirement. This requirement is implied by, but does not imply belief independence.

Definition IV.11. A Bayesian equilibrium of game form G for every type space, σ^*, are *consistent* if for all type spaces $\mathcal{T}, \widetilde{\mathcal{T}} \in \Upsilon$ such that:

(i) for every $i \in I$: $\widetilde{T}_i \subseteq T_i$ (where $\widetilde{T}_i$ corresponds to $\widetilde{\mathcal{T}}$ and T_i corresponds to $\mathcal{T}$);

(ii) for every $i \in I$ and every $t_i \in T_i$: $\tilde{u}_i(t_i) = u_i(t_i)$ and $\tilde{\pi}_i(t_i) = \pi_i(t_i)$ (where $\tilde{u}_i, \tilde{\pi}_i$ correspond to $\widetilde{\mathcal{T}}$, and u_i, π_i correspond to $\mathcal{T}$),

we have for every $i \in I$ and every $t_i \in T_i$:

(iii) $\sigma^*(\widetilde{\mathcal{T}}, t_i) = \sigma^*(\mathcal{T}, t_i)$.

Observe that the type t_i referred to in item (iii) of Definition V.4 has the same utility function and hierarchy of beliefs in type space $\mathcal{T}$ and in type space $\widetilde{\mathcal{T}}$. Therefore, the consistency requirement is implied by the assumption that an agent's equilibrium choices only depend on that agent's utility function and that agent's hierarchy of beliefs. This assumption seems reasonable because the type space, as opposed to the utility function and the hierarchy of beliefs, is really only a construction by the modeler, and not necessarily a construction that the agent is aware of. We don't explicitly formulate the stronger assumption that equilibrium choices should only depend on agents' utility functions and hierarchies of beliefs, but instead work with the weaker consistency requirement, because the consistency requirement is easier to formulate,

and is sufficient for our purposes. Our results would also go through if we made the more demanding assumption for equilibria.

4.5 A Game Form that Interim Pareto Dominates Random Dictatorship

The first main result of this paper examines interim Pareto dominance, while the second main result concerns ex post Pareto dominance. The first result says that for every $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$ and $p < 1$ for all $i \in I$ there are a game form, and a Bayesian equilibrium of this game form for every type space, that interim Pareto dominate p random dictatorship. We refer to the dominating game form as *p-random dictatorship with compromise.*

Definition IV.12. For every $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$ the following game form is called a *p-random dictatorship with compromise.*

(i) for every $i \in I$:

$$S_i = 2^A \times A,$$

where 2^A is the set of all non-empty subsets of A;

(ii) If $a_i = a$ for some $a \in A$ and all $i \in I$, then:

$$x(s,a) = 1$$

(iii) If $a_i \neq a_j$ for some $i, j \in I$, but $\bigcap_{i\in I} \mathcal{A}_i \neq \emptyset$, then there is some $a \in \bigcap_{i\in I} \mathcal{A}_i$ such that

$$x(s,a) = 1.$$

(iv) If $a_i \neq a_j$ for some $i, j \in I$, and $\bigcap_{i \in I} \mathcal{A}_i = \emptyset$, then for all $a \in A$:

$$x(s, a) = \sum_{\{i \in I : a_i = a\}} p_i.$$

In words, this game form offers each agent i the opportunity to nominate one "preferred" alternative, a_i, and also a set $\mathcal{A}_i$ of "acceptable" alternatives. If all voters nominate the same preferred alternative, then that alternative is chosen with probability 1. If voters' preferred alternatives differ, but there is at least one alternative that all voters include in their set of acceptable alternatives, then one of the commonly acceptable alternatives is chosen with probability 1. Otherwise, the mechanism reverts to random dictatorship. We refer to this game form as *p random dictatorship with compromise* because it offers agents the opportunity to compromise on a mutually acceptable alternative in place of p random dictatorship.[12]

One Bayesian equilibrium of this game form is that all agents always choose a_i to be their most preferred alternative, and set $\mathcal{A}_i = \{a_i\}$. In this equilibrium, the possibility of a compromise is not used by either agent. This is an equilibrium because neither agent can unilaterally force a compromise. Any deviation that unilaterally alters the set of acceptable alternatives has no effect. However, the next proposition shows that *p-random dictatorship with compromise* also has a Bayesian equilibrium for all type spaces that interim Pareto dominates random dictatorship. We also show that this equilibrium respects positive and negative unanimity, to clarify that our result does indeed result from weakening the belief independence requirement, and not weakening any other property listed in Proposition IV.9.

[12]This game form was inspired by the idea of *Approval Voting* (see Brams and Fishburn, 2007), which, like our game form, allows voters to indicate "acceptable" alternatives. However, in approval voting the alternative that the largest number of agents regards as acceptable is selected, whereas our game form requires unanimity. Moreover, our game form uses random dictatorship as a fallback, whereas approval voting does not have any such fallback. When p is the uniform distribution, the game form that we consider is almost identical to the *Full Consensus or Random Ballot Fall-Back* game form that Heitzig and Simmons (2010) introduced. Heitzig and Simmons require the set $\mathcal{A}_i$ to be a singleton.

Proposition IV.13. *For every $p \in [0,1]^n$ such that $\sum_{i \in I} p_i = 1$ and $p_i < 1$ for all $i \in I$, p-random dictatorship with compromise has a consistent equilibrium for all type spaces σ^* that interim Pareto dominates p-random dictatorship and that satisfies positive and negative unanimity.*

The main difficulty in the proof below is not so much showing interim Pareto dominance, but proving the existence of a consistent equilibrium. The argument in the proof below can be used to show the existence of consistent Bayesian equilibria for all type spaces of arbitrary finite games.

Proof. We construct the equilibrium σ^*. To begin with we restrict attention to strategies such that $a_i = b(u_i(t_i))$ and $w(u_i(t_i)) \notin \mathcal{A}_i$. This restriction of the strategy space is innocuous, because any strategy that does not satisfy this restriction is weakly dominated by a strategy that does satisfy it. This restriction implies that positive and negative unanimity will automatically be satisfied.

We now proceed inductively. We begin by considering type spaces $\mathcal{T}$ where for every $i \in I$ the set T_i has exactly one element. In such type spaces it is common belief among the agents that agent i has utility function $u_i(t_i)$. We distinguish two cases. The first is that there is some alternative $a \in A$ such that for all $i \in I$ we have:

$$u_i(t_i, a) > \sum_{j \in I} p_j u_i(t_i, b(u_j(t_j))). \tag{4.10}$$

Observe that the assumption $p < 1$ for all $i \in I$ implies that some such type spaces exist. For such type spaces the strategies are:

$$\sigma_i(\mathcal{T}, t_i) = (\{b(u_i(t_i)), a\}, b(u_i(t_i))) \tag{4.11}$$

for $i \in I$. Note that these strategies constitute a Nash equilibrium of the complete information game in which agents' preferences are common knowledge, and that the

outcome a strictly Pareto-dominates the outcome under random dictatorship. For all other type spaces with just a single element for each player the strategies are:

$$\sigma_i(\mathcal{T}, t_i) = (\{b(u_i(t_i))\}, b(u_i(t_i))) \tag{4.12}$$

Note that these strategies constitute a Nash equilibrium of the complete information game in which agents' preferences are common knowledge, and that the outcome is exactly the same as under random dictatorship.

Now suppose we had constructed the equilibrium for all type spaces $\mathcal{T}$ in which all the sets T_i have at most k elements. We first extend the construction to all type spaces $\mathcal{T}$ in which T_1 has at most $k+1$ elements and for $j > 1$ the set T_j has at most k elements. Then we extend the construction to all type spaces $\mathcal{T}$ in which T_1 and T_2 have at most $k+1$ elements and for $j > 2$ the set T_j has at most k elements. The construction can then inductively continued until it is extended to all type spaces $\mathcal{T}$ in which all the sets T_i have at most $k+1$ elements.

Suppose first that we are considering a type space $\mathcal{T}$ in which T_i has at most $k+1$ elements and for $j > 1$ the set T_j has at most k elements. Consider all type spaces $\widetilde{\mathcal{T}}$ that are contained in $\mathcal{T}$, i.e. for which conditions (i) and (ii) of Definition V.4 hold, and such that at least for one agent the type set has fewer elements than in $\mathcal{T}$. For such type spaces we define for every $i \in I$ and every $t_i \in \widetilde{T}_i$:

$$\sigma_i\left(\mathcal{T}, t_i\right) = \sigma_i\left(\widetilde{\mathcal{T}}, t_i\right). \tag{4.13}$$

By the inductive hypothesis the right hand side of this equation has already been defined. Observe that this is well-defined. If a type t_i of player i is contained in player i's type set in two different type spaces $\widetilde{\mathcal{T}}$ and $\widehat{\mathcal{T}}$ that are contained in $\mathcal{T}$ in the sense of Definition V.4, then the intersection of these type spaces is also a type space, and by consistency the same strategy is assigned to type t_i in $\widetilde{\mathcal{T}}$ and in $\widehat{\mathcal{T}}$.

If the previous step defines the equilibrium strategy for all types in $\mathcal{T}$, then the inductive step is completed. Otherwise, it remains to define strategies for types t_i that are not contained in any type set of a type space that is a subspace of $\mathcal{T}$. We consider the strategic game in which each such type is a separate player, and expected utilities are calculated keeping the strategies of types that have already been dealt with in the previous paragraph fixed, and using each type's subjective beliefs to calculate that type's expected payoff. This strategic game has a Nash equilibrium in mixed strategies. We define for each type t_i that still has to be dealt with the strategy $\sigma_i(\mathcal{T}, t_i)$ to be type t_i's equilibrium strategy.

By construction these strategies satisfy the consistency requirement. Also, they are by construction interim Bayesian equilibria: For types in typesets that correspond to a smaller type space the Bayesian equilibrium property carries over from the smaller type space. For all other types, their choices maximize expected utility by construction.

We extend the construction to all type spaces $\mathcal{T}$ in which T_1 and T_2 have at most $k+1$ and for $i > 2$ the set T_i has at most k elements in the same way as we extended it to all type spaces $\mathcal{T}$ in which T_1 has at most $k+1$ elements and for $i > 1$ the set T_i has at most k elements.

To conclude the proof we note that this equilibrium interim Pareto dominates random dictatorship. First, we note that no type can have lower expected utility than under random dictatorship. This is because each type can guarantee themselves an outcome that is at least as good as the random dictatorship outcome by choosing $\mathcal{A}_i = \{b(u_i(t_i))\}$. Second, each type's expected utility is increased on type spaces in which each player's type set has just a single element, and for which inequality (4.10) holds. □

4.6 No Game Form Ex Post Pareto Dominates Random Dictatorship

Proposition IV.14. *For every $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$ there is no game form G that has a consistent equilibrium for all type spaces σ^* that ex post Pareto dominates p-random dictatorship.*

Proof. Indirect. Suppose for some $p \in [0,1]^n$ such that $\sum_{i\in I} p_i = 1$ there were a game form G and an equilibrium of G for all type spaces σ^* that ex post Pareto dominate p-random dictatorship. For the outcome resulting from G and σ^* to be different from p-random dictatorship, there must be some $\widehat{\mathcal{T}} \in \Upsilon$, $\hat{t} \in \widehat{T}$, and $\hat{a} \in A$ such that:

$$\sum_{s\in S} x(s,\hat{a}) \cdot \prod_{i\in I} \sigma_i^*(\widehat{\mathcal{T}}, \hat{t}_i, s_i) < \sum_{\left\{i\in I : b(u_i(\hat{t}_i))=\hat{a}\right\}} p_i. \tag{4.14}$$

That is, alternative $\hat{a}$ is chosen with a probability that is strictly smaller than the probability with which it is chosen under random dictatorship. Let $\widehat{I}$ be the set $\left\{i \in I : b(u_i(\hat{t}_i)) = \hat{a}\right\}$, and notice that this set must be non-empty for (4.14) to hold. To complete the proof we construct a new type space $\widetilde{\mathcal{T}}$, and infer from (4.14) that in this type space there is a type vector such that the outcome of p-random dictatorship conditional on this type vector is strictly preferred by the type of one of the players in $\widehat{I}$ to the outcome in G resulting from the equilibrium σ^*. Therefore, G and σ^* do not ex post Pareto-dominate p-random dictatorship.

The type sets in $\widetilde{\mathcal{T}}$ are given by: $\widetilde{T}_i = \widehat{T}_i$ for all $i \in \widehat{I}$, and $\widetilde{T}_i = \widehat{T}_i \cup \{\tilde{t}_i\}$ for all $i \notin \widehat{I}$. For all $i \in I$ The types in $\widehat{T}_i$ have the same utility functions and beliefs in $\widetilde{\mathcal{T}}$ as in $\widehat{\mathcal{T}}$. For all $i \notin \widehat{I}$ type $\tilde{t}_i$'s beliefs are given by:

$$\pi_i(\tilde{t}_i)\left[\left(\left(\hat{t}_j\right)_{j\in\widehat{I}}, \left(\tilde{t}_j\right)_{\substack{j\notin\widehat{I}\\ j<i}}, \left(\hat{t}_j\right)_{\substack{j\notin\widehat{I}\\ j>i}}\right)\right] = 1, \tag{4.15}$$

and type $\tilde{t}_i$'s utility function is:

$$\tilde{u}_i(\tilde{t}_i, a) = \begin{cases} 1 & \text{if } a = \tilde{a}; \\ 1 - \varepsilon_a & \text{if } a \notin \{\hat{a}, \tilde{a}\}; \\ 0 & \text{if } a = \hat{a}; \end{cases} \tag{4.16}$$

where $\tilde{a}$ denotes the second most preferred alternative of some player k's type $\hat{t}_k$, where $k \in \widehat{I}$. We assume that $0 < \varepsilon_a < \bar{\varepsilon}$ for all $a \notin \{\hat{a}, \tilde{a}\}$ for some $\bar{\varepsilon} \in (0, 1)$, and that $a, a' \notin \{\hat{a}, \tilde{a}\}$ and $a \neq a'$ implies $\varepsilon_a \neq \varepsilon_{a'}$. This assumption ensures that the utility functions satisfy the condition of no indifferences. Moreover, by letting $\bar{\varepsilon}$ tend to zero, we can ensure that all ε_a tend to zero, which is the case that we shall focus on.

We now show that for $\bar{\varepsilon}$ sufficiently small at type vector $((\hat{t}_i)_{i \in \hat{I}}, (\tilde{t}_i)_{i \notin \hat{I}})$ the alternatives other than $\hat{a}$ are in equilibrium σ^* chosen with a total probability that is larger than $1 - \sum_{i \in \hat{I}} p_i$. Note that the proof of Proposition V.13 is concluded once this assertion is established. This is because random dictatorship gives player k's type $\hat{t}_k$ his top alternative $\hat{a}$ with probability $\sum_{i \in \hat{I}} p_i$, and type $\hat{t}_k$'s second most preferred alternative $\tilde{a}$ with probability $1 - \sum_{i \in \hat{I}} p_i$. By contrast, G and σ^* yield $\hat{a}$ with probability less than $\sum_{i \in \hat{I}} p_i$, and some other alternative, not necessarily type $\hat{t}_k$'s second most preferred alternative, with a probability larger than $1 - \sum_{i \in \hat{I}} p_i$. Therefore, type $\hat{t}_k$ strictly prefers random dictatorship.

Consider the player $i \notin \widehat{I}$ for whom i is smallest. We denote this player by $i1$. This player, when type $\tilde{t}_{i1}$, expects with probability 1 that the other players' type vector is $\hat{t}_{-i1}$. Because σ^* is consistent, type $\tilde{t}_{i1}$ expects the types $\hat{t}_{-i1}$ to choose the same in $\widetilde{T}$ as in $\widehat{T}$. By the assumption of the indirect proof, type $\hat{t}_{i1}$ has a strategy available that yields alternatives other than $\hat{a}$ with probability of more than $1 - \sum_{i \in \hat{I}} p_i$. Type $\tilde{t}_{i1}$ will not necessarily choose the same strategy as type $\hat{t}_{i1}$. But,

for small enough $\bar{\varepsilon}$, only a strategy that yields an alternative other than $\hat{a}$ with some probability $\tilde{p} > 1 - \sum_{i \in \widehat{I}} p_i$ can be optimal. Choosing such a strategy yields for type $\tilde{t}_{i1}$ expected payoff greater than $\tilde{p}(1-\bar{\varepsilon}) > \left(1 - \sum_{i \in \widehat{I}} p_i\right)(1-\bar{\varepsilon})$ whereas any other pure strategy yields a a payoff that is no more than $1 - \sum_{i \in \widehat{I}} p_i < \tilde{p}$. For small enough $\bar{\varepsilon}$ the former expected payoff is larger than the latter.

Now consider the player $i \notin \widehat{I}$ for whom i is second smallest. We denote this player by $i2$. This player, when type $\tilde{t}_{i2}$, expects with probability 1 the other players' types to be $\hat{t}_{-i2}$ except for player $i1$ whom $i2$ expects with probability 1 to be type $\tilde{t}_{i1}$. By the step of the previous paragraph, if $\tilde{t}_{i2}$ chose the same strategy as $\hat{t}_{i2}$ does in equilibrium, $\tilde{t}_{i2}$ would expect an outcome other than $\hat{a}$ with probability larger than $1 - \sum_{i \in \widehat{I}} p_i$. He might choose in equilibrium some other strategy, but, for small enough $\bar{\varepsilon}$, he will never make a choice that yields an outcome other than $\hat{a}$ with a probability that is not larger than $1 - \sum_{i \in \widehat{I}} p_i$.

The step of the previous paragraph can be iterated until we arrive at the player $i \notin \widehat{I}$ for whom i is largest. We denote this player by $i(n-1)$. This player expects the other players to be of type $\tilde{t}_{-(i(n-1))}$ except for types $i \in \widehat{I}$, whom this player expects to be of type $\hat{t}_i$. By the same argument used in the previous two paragraphs, type $\tilde{t}_{i(n-1)}$ chooses in equilibrium a strategy that he expects to yield an outcome other than $\hat{a}$ with probability larger than $1 - \sum_{i \in \widehat{I}} p_i$. But at type vector $((\hat{t}_i)_{i \in \hat{I}}, (\tilde{t}_i)_{i \notin \widehat{I}})$ this type has correct expectations, and therefore at this type vector the equilibrium strategies do indeed yield an outcome other than $\hat{a}$ with probability larger than $1 - \sum_{i \in \widehat{I}} p_i$. As explained above, this concludes the proof. □

4.7 Conclusion

Gibbard and Satterthwaite's theorem, and Hylland's version of this theorem in a cardinal utility setting, are central results of voting theory. We have argued that

the insistence of the theorem on belief independent strategy choices may be overly restrictive if a mechanism designer is considered who is concerned with Pareto improvements. Such a mechanism designer can find voting schemes that are superior to random dictatorship if agents' choices are allowed to depend on their beliefs. Whatever those beliefs are, the outcomes will be at least as good as under random dictatorship, and sometimes better. Such an improvement is only possible if agents' subjective beliefs are accepted, and an interim perspective is adopted. From an ex post perspective, such unambiguous improvements are not possible.

An important problem left open by our paper is the characterization of voting rules than are not dominated in one of the senses considered in this paper. In chapter 4 the analogous question is investigated for public goods mechanisms. Chapter 4 proves for one particular mechanism that it is not dominated. That chapter shows the subtleties of this problem. Other related work is by Azrieli and Kim (2011) who consider interim efficient voting rules for 2 alternatives, restricting attention to independent types, and Börgers and Postl (2009) who describe ex ante efficient voting schemes over three alternatives, but who only consider a very restricted type space with, in particular, independent types.

Another important step is the investigation of robust implementation as opposed to robust mechanism design. Implementation, unlike mechanism design, considers *all* equilibria of a given game form. One might ask whether there are mechanisms such that *all* equilibria on all type spaces dominate random dictatorship. We leave this question for future research.

References

Yaron Azrieli and Semin Kim (2011), Pareto Efficiency and Weighted Majority Rules, unpublished, Ohio State University.

Salvador Barberà (2010), Strategy-proof Social Choice, Chapter 25 in: K. J. Arrow, A. K. Sen and K. Suzumura (eds.), *Handbook of Social Choice and Welfare*, Amsterdam: North-Holland.

Jean-Marie Blin and Mark Satterthwaite (1977), On Preferences, Beliefs, and Manipulation within Voting Situations, *Econometrica* 45, 881-888.

Dirk Bergemann and Stephen Morris (2003), Robust Mechanism Design, Cowles Foundation Discussion Paper No. 1421.

Dirk Bergemann and Stephen Morris (2005), Robust Mechanism Design, *Econometrica* 73, 1771-1813.

Tilman Börgers (1991), Undominated Strategies and Coordination in Normalform Games, *Social Choice and Welfare* 8, 65-78.

Tilman Börgers and Peter Postl (2009), Efficient Compromising, *Journal of Economic Theory* 155, 2057-2076.

Steven Brams and Peter Fishburn (2007), *Approval Voting*, second edition, Heidelberg etc.: Springer.

Bhaskar Dutta, Hans Peters, and Arunava Sen (2007), Strategy-Proof Cardinal Decision Schemes, *Social Choice and Welfare* 28, 163-179.

Bhaskar Dutta, Hans Peters, and Arunava Sen (2008), Strategy-Proof Cardinal Decision Schemes (Erratum), *Social Choice and Welfare* 30, 701-702.

Drew Fudenberg and Jean Tirole (1991), *Game Theory*, Cambridge and London: The MIT

Press.

Alan Gibbard (1973), Manipulation of Voting Schemes: A General Result, *Econometrica* 41, 587-602.

Jobst Heitzig and Forrest Simmons (2010), Some Chance for Consensus: Voting Methods For Which Consensus is an Equilibrium, *Social Choice and Welfare*, forthcoming.

Aanund Hylland (1980), Strategy Proofness of Voting Procedures with Lotteries as Outcomes and Infinite Sets of Strategies, mimeo., University of Oslo, Institute of Economics.

Bengt Holmström and Roger Myerson (1983), Efficient and Durable Decision Rules with Incomplete Information, *Econometrica* 51, 1799-1819.

Shasikanta Nandeibam (2004), The Structure of Decision Schemes with von Neumann Morgenstern Preferences, discussion paper, University of Bath.

Mark Satterthwaite (1975), Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions, *Journal of Economic Theory* 1975, 187-217.

CHAPTER V

Robust Mechanism Design and Dominant Strategy Voting Rules: The Relative Utilitarianism Case

5.1 Introduction

This paper extends the analysis of voting mechanisms in chapter 4, to the case in which a mechanism designer considers the relative efficiency of mechanisms' outcomes from the perspective of relative utilitarianism. Both papers analyze efficiency without making any assumption regarding voters' knowledge about each other. Chapter 4 analyzed voting rules in terms of Pareto efficiency. It found that on full domains, the only dominant strategy voting rules are random dictatorships, and that the designer of a voting rule can achieve Pareto improvements over random dictatorship by choosing rules in which voters' behavior can depend on their beliefs, although the result only holds for voters' interim expected utilities, not for their ex post expected utilities.

This paper focuses on the two agent environment and extends the results to the case where a mechanism designer considers an outcome efficient if it maximizes the sum of agents' utilities, after agents' utilities have been normalized. This "relative utilitarian" welfare function was axiomatized by Dhillon (1998) and Dhillon and Mertens (1999) for social choice problems.

The paper is organized as follows. Section 5.2 describes the voting problem, which

is largely identical to chapter 4 except for the assumption that there are two agents, and that utilities are normalized in a way that permits a utilitarian approach. Section 5.3 introduces our notions of efficiency and shows that the result, in chapter 4, that random dictatorship is interim Pareto dominated immediately implies that random dictatorship is also interim dominated from the relative utilitarianism perspective. Section 5.4 extends the second main result from that paper by showing that no game form ex post dominates random dictatorship when efficiency is defined in terms of relative utilitarianism. Section 5.5 concludes.

5.2 The Voting Problem

There are two agents: $i \in \{1, 2\}$. The agents have to choose one alternative from a finite set A of alternatives. We assume that A has at least three elements. The set of all probability distributions over A is $\Delta(A)$, where for $\delta \in \Delta(A)$ we denote by $\delta(a) \in [0, 1]$ the probability that δ assigns to alternative a. The two agents are commonly known to be expected utility maximizers. We denote agent i's von Neumann Morgenstern utility function by $u_i : A \to \mathbb{R}$. We assume that each agent's von Neumann Morgenstern utility function is normalized such that $\min_{a \in A} u_i(a) = 0$ and $\max_{a \in A} u_i(a) = 1$. We also assume that $a \neq b \Rightarrow u_i(a) \neq u_i(b)$, i.e., there are no indifferences. We define the expected utility for probability distributions $\delta \in \Delta(A)$ by $u_i(\delta) = \sum_{a \in A} u_i(a) \cdot \delta(a)$.

A mechanism designer has a ranking of the alternatives in A that may depend on the agents' utility functions. We shall be more specific about the designer's objectives later. The mechanism designer does not know the agents' utility functions, nor does she know what the agents believe about each other. To implement an outcome that potentially depends on the agents' utility functions the mechanism designer asks the agents to play a *game form.*

Definition V.1. A *game form* $G = (S_1, S_2, x)$ consists of:

(i) a non-empty finite strategy set S_i for each agent $i \in \{1, 2\}$;

We define: $S \equiv S_1 \times S_2$.

(ii) an outcome function $x : S \to \Delta(A)$.

The set S_i is the set of (pure) strategies available to agent i in the game form G. We focus on finite sets of pure strategies, while allowing mixed strategies, to ease exposition. Our results also hold when the sets S_i of pure strategies are allowed to be infinite. The function x assigns to every combination of pure strategies s the, potentially stochastic, outcome $x(s)$ that is implemented when agents choose that combination of pure strategies. We write $x(s, a)$ for the probability that $x(s)$ assigns to alternative a.

Once the mechanism designer has announced a game form, the two agents choose simultaneously and independently their strategies. Because the agents don't necessarily know each others' utility functions or beliefs, this game may be a game of incomplete information. A hypothesis about the agents' utility functions and their beliefs about each other can be described by specifying a *type space.*

Definition V.2. A *type space* $\mathcal{T} = (T_1, T_2, \pi_1, \pi_2, u_1, u_2)$ consists for each $i \in \{1, 2\}$ of:

(i) a nonempty, finite set T_i of types;

We write $\Delta(T_i)$ for the set of all probability distributions over T_i.

(ii) a belief function $\pi_i : T_i \to \Delta(T_j)$ (where $j \neq i$);

(iii) a utility function $u_i : T_i \times A \to [0, 1]$.

We write $\pi_i(t_i, t_j)$ for the probability that type i assigns to player j being type t_j (where $j \neq i$). We write $u_i(t_i, a)$ for the utility that $u_i(t_i)$ assigns to a.[1] The utility function satisfies for both $i \in \{1, 2\}$ and all $t_i \in T_i$ the assumptions introduced earlier:

(a) $\min_{a \in A} u_i(t_i, a) = 0$ and $\max_{a \in A} u_i(t_i, a) = 1$;

(b) $u_i(t_i, a) \neq u_i(t_i, b)$ whenever $a \neq b$.

The set T_i is the set of types of agent i. Agent i privately observes his type. The function π_i describes for every type $t_i \in T_i$ the beliefs that agent i has about the other agents' types when agent i himself is of type t_i. We write $\pi_i(t_i, t_{-i})$ for the probability that type t_i assigns to the other players types being t_{-i}. The function $u_i(t_i)$ describes player i's utility when i is of type t_i. We write $u_i(t_i, a)$ for the utility that $u_i(t_i)$ assigns to alternative a. The utility functions $u_i(t_i)$ satisfy the assumption that we introduced earlier that there are no indifferences.

In Definition V.2 beliefs are subjective. There may or may not be a common prior for a particular type space. Different agents' beliefs may be incompatible with each other in the sense that one agent may attach probability one to an event to which another agent attaches probability zero. Observe also that we assume type spaces to be finite. We thus avoid technical difficulties associated with infinite type spaces.

We assume that the mechanism designer has no knowledge of the agents' utility functions or their beliefs. Therefore, the mechanism designer regards all type spaces as possible descriptions of the environment in which agents find themselves. We denote the set of all type spaces by Υ.

The mechanism designer proposes to agents how they might play the game. He might propose to agents to randomize. For $i = 1, 2$ we denote by $\Delta(S_i)$ the set of all probability distributions on S_i. For the agents to accept the mechanism designer's

[1] Observe that we suppress in the notation the dependence of π_i and u_i on the type space $\mathcal{T}$. We are not aware of any confusion that might arise from this simplification of our notation.

proposal, he must propose a *Bayesian equilibrium*. Because the mechanism designer does not know the true type space, he has to propose a *Bayesian equilibrium for every type space*.

Definition V.3. A *Bayesian equilibrium of game form G for every type space* is a pair (σ_1^*, σ_2^*) such that for every $i \in \{1, 2\}$:

(i) σ_i^* is a family of functions $(\sigma_i^*(\mathcal{T}))_{\mathcal{T} \in \Upsilon}$ where for every $\mathcal{T} \in \Upsilon$ the function $\sigma_i^*(\mathcal{T})$ maps the type space T_i corresponding to $\mathcal{T}$ into $\Delta(S_i)$.

We write $\sigma_i^*(\mathcal{T}, t_i)$ for the mixed strategy assigned to $t_i \in T_i$, and $\sigma_i^*(\mathcal{T}, t_i, s_i)$ for the probability that this mixed strategy assigns to $s_i \in S_i$.

(ii) $\sigma_i^*(\mathcal{T}, t_i)$ maximizes the expected utility of type t_i among all mixed strategies in $\Delta(S_i)$, where expected utility for any mixed strategy $\sigma_i^* \in \Delta(S_i)$ is:

$$\sum_{t_j \in T_j} \sum_{s_1 \in S_1, s_2 \in S_2} \left(u_i(t_i, x(s_1, s_2)) \cdot \sigma_i^*(\mathcal{T}, t_i, s_i) \cdot \sigma_j^*(\mathcal{T}, t_j, s_j) \cdot \pi(t_i, t_j) \right), \qquad (5.1)$$

where $j \neq i$.

The Bayesian equilibria that the mechanism designer proposes need to satisfy a *consistency* requirement.

Definition V.4. A Bayesian equilibrium of game form G for every type space, (σ_1^*, σ_2^*), is *consistent* if for all type spaces $\mathcal{T}, \widetilde{\mathcal{T}} \in \Upsilon$ such that:

(i) for every $i \in \{1, 2\}$: $\widetilde{T}_i \subseteq T_i$ (where $\widetilde{T}_i$ corresponds to $\widetilde{\mathcal{T}}$ and T_i corresponds to $\mathcal{T}$);

(ii) for every $i \in \{1, 2\}$ and every $t_i \in T_i$: $\tilde{u}_i(t_i) = u_i(t_i)$ and $\tilde{\pi}(t_i) = \pi(t_i)$ (where $\tilde{u}_i, \tilde{\pi}_i$ correspond to $\widetilde{\mathcal{T}}$, and u_i, π_i correspond to $\mathcal{T}$),

we have for every $i \in \{1, 2\}$ and every $t_i \in T_i$:

(iii) $\sigma^*(\widetilde{\mathcal{T}}, t_i) = \sigma^*(\mathcal{T}, t_i)$.

Observe that the type t_i referred to in item (iii) of Definition V.4 has the same utility function and hierarchy of beliefs in type space $\mathcal{T}$ and in type space $\widetilde{\mathcal{T}}$. Therefore, the consistency requirement is implied by the assumption that an agent's equilibrium choices should only depend on that agent's utility function and that agent's hierarchy of beliefs. This assumption seems reasonable because the type space, as opposed to the utility function and the hierarchy of beliefs, is really only a construction by the modeler, and not necessarily a construction that the agent is aware of. We don't explicitly formulate the stronger assumption that equilibrium choices should only depend on agents' utility functions and hierarchies of beliefs, but instead work with the weaker consistency requirement, because the consistency requirement is easier to formulate, and is sufficient for our purposes. Our results would also go through if we made the more demanding assumption for equilibria.

We postulate a mechanism designer who seeks to further the utility of the agents rather than his own utility. In the companion paper, Börgers and Smith **?**, we assume that the mechanism designer seeks to achieve a Pareto efficient decision. In this paper, we assume that the mechanism designer seeks to maximize the welfare function: $u_1(a) + u_2(a)$. Because we have normalized utilities, this corresponds to the "relative utilitarian" welfare function axiomatized by Dhillon (1998) and Dhillon and Mertens (1999).

When evaluating the utility of the two agents for a realized type combination (t_1, t_2) the mechanism designer can either only consider the outcomes that result from the mixed strategies prescribed for these two types, or she may consider the expected utilities of these two types, based on the types' own subjective beliefs. In other words, the mechanism designer may adopt an *ex post* or an *interim* perspective when evaluating agents' utilities. The interim perspective respects agents' own perception of their environment. From this perspective, the ex post perspective has a paternalistic

flavor. On the other hand, for example when agents' beliefs are incompatible with each other, the mechanism designer may be justified in discarding agents' beliefs, on the basis that at least some of them have to be wrong, as agents themselves will discover at some point. Thus neither the interim nor the ex post perspective are clearly preferable. We pursue both perspectives in this paper.

The considerations of the preceding two paragraphs lead to four possible formalizations of the mechanism designer's objectives. We present these in the four definitions that follow below. None of these definitions attributes a prior over type spaces in Υ or over types in each type space to the mechanism designer. Instead, we work with a dominance notion, that is prior free. Whatever the mechanism designer's prior is, if he has one, he will never choose a dominated game form in the sense described in the four definitions below.

Definition V.5. The game form G with the consistent Bayesian equilibrium for all type spaces (σ_1^*, σ_2^*) *ex post Pareto dominates* the game form $\tilde{G}$ with the consistent Bayesian equilibrium for all type spaces $(\tilde{\sigma}_1^*, \tilde{\sigma}_2^*)$ if for all $i \in \{1,2\}$, $\mathcal{T} \in \Upsilon$, and $(t_1, t_2) \in T_1 \times T_2$:

$$\sum_{s_1 \in S_1, s_2 \in S_2} u_i(t_i, x(s_1, s_2)) \cdot \sigma_1^*(\mathcal{T}, t_1, s_1) \cdot \sigma_2^*(\mathcal{T}, t_2, s_2) \geq$$
$$\sum_{s_1 \in \tilde{S}_1, s_2 \in \tilde{S}_2} u_i(t_i, \tilde{x}(s_1, s_2)) \cdot \tilde{\sigma}_1^*(\mathcal{T}, t_1, s_1) \cdot \tilde{\sigma}_2^*(\mathcal{T}, t_2, s_2), \tag{5.2}$$

with strict inequality for at least one $i \in \{1,2\}$, $\mathcal{T} \in \Upsilon$, and $(t_1, t_2) \in T_1 \times T_2$. A direct mechanism that is not ex post Pareto dominated will be called *ex post Pareto undominated.*

Definition V.6. The game form G with the consistent Bayesian equilibrium for all type spaces (σ_1^*, σ_2^*) *ex post utilitarian*[2] *dominates* the game form $\tilde{G}$ with the consistent Bayesian equilibrium for all type spaces $(\tilde{\sigma}_1^*, \tilde{\sigma}_2^*)$ if for all $\mathcal{T} \in \Upsilon$, and

[2]For simplicity, we use "utilitarian" rather than the more clumsy "relative utilitarian."

$(t_1, t_2) \in T_1 \times T_2$:

$$\sum_{i\in\{1,2\}} \sum_{s_1\in S_1, s_2\in S_2} u_i(t_i, x(s_1, s_2)) \cdot \sigma_1^*(\mathcal{T}, t_1, s_1) \cdot \sigma_2^*(\mathcal{T}, t_2, s_2) \geq$$
$$\sum_{i\in\{1,2\}} \sum_{s_1\in \tilde{S}_1, s_2\in \tilde{S}_2} u_i(t_i, \tilde{x}(s_1, s_2)) \cdot \tilde{\sigma}_1^*(\mathcal{T}, t_1, s_1) \cdot \tilde{\sigma}_2^*(\mathcal{T}, t_2, s_2), \tag{5.3}$$

with strict inequality for at least one $\mathcal{T} \in \Upsilon$, and $(t_1, t_2) \in T_1 \times T_2$. A direct mechanism that is not ex post utilitarian dominated will be called *ex post utilitarian undominated.*

Note that if game form G with the consistent Bayesian equilibrium for all type spaces (σ_1^*, σ_2^*) ex post utilitarian dominates game form $\tilde{G}$ with the consistent Bayesian equilibrium for all type spaces $(\tilde{\sigma}_1^*, \tilde{\sigma}_2^*)$ if the former ex post Pareto dominates the latter.

Definition V.7. The game form G with the consistent Bayesian equilibrium for all type spaces (σ_1^*, σ_2^*) *interim Pareto dominates* the game form $\tilde{G}$ with the consistent Bayesian equilibrium for all type spaces $(\tilde{\sigma}_1^*, \tilde{\sigma}_2^*)$ if for all $i, j \in \{1, 2\}$ with $i \neq j$, $\mathcal{T} \in \Upsilon$, and $t_i \in T_i$:

$$\sum_{t_j\in T_j} \pi_i(t_i, t_j) \sum_{s_1\in S_1, s_2\in S_2} u_i(t_i, x(s_1, s_2)) \cdot \sigma_1^*(\mathcal{T}, t_1, s_1) \cdot \sigma_2^*(\mathcal{T}, t_2[s_2) \geq$$
$$\sum_{t_j\in T_j} \pi_i(t_i, t_j) \sum_{s_1\in \tilde{S}_1, s_2\in \tilde{S}_2} u_i(t_i, \tilde{x}(s_1, s_2)) \cdot \tilde{\sigma}_1^*(\mathcal{T}, t_1, s_1) \cdot \tilde{\sigma}_2^*(\mathcal{T}, t_2, s_2), \tag{5.4}$$

with strict inequality for at least one $i, j \in \{1, 2\}$ with $i \neq j$, $\mathcal{T} \in \Upsilon$, and $t_i \in T_i$. A direct mechanism that is not interim Pareto dominated will be called *interim Pareto undominated.*

Definition V.8. The game form G with the consistent Bayesian equilibrium for all type spaces (σ_1^*, σ_2^*) *interim utilitarian dominates* the game form $\tilde{G}$ with the consistent

Bayesian equilibrium for all type spaces $(\tilde{\sigma}_1^*, \tilde{\sigma}_2^*)$ if for all $\mathcal{T} \in \Upsilon$ and $(t_1, t_2) \in T_1 \times T_2$:

$$\sum_{i\in\{1,2\}} \sum_{t_j \in T_j} \pi_i(t_i, t_j) \sum_{s_1 \in S_1, s_2 \in S_2} u_i(t_i, x(s_1, s_2)) \cdot \sigma_1^*(\mathcal{T}, t_1, s_1) \cdot \sigma_2^*(\mathcal{T}, t_2, s_2) \geq$$
$$\sum_{i\in\{1,2\}} \sum_{t_j \in T_j} \pi_i(t_i, t_j) \sum_{s_1 \in \tilde{S}_1, s_2 \in \tilde{S}_2} u_i(t_i, \tilde{x}(s_1, s_2)) \cdot \tilde{\sigma}_1^*(\mathcal{T}, t_1, s_1) \cdot \tilde{\sigma}_2^*(\mathcal{T}, t_2, s_2), \quad (5.5)$$

with strict inequality for at least one $\mathcal{T} \in \Upsilon$ and $(t_1, t_2) \in T_1 \times T_2$. A direct mechanism that is not interim utilitarian dominated will be called *interim utilitarian undominated.*

Note that if game form G with the consistent Bayesian equilibrium for all type spaces (σ_1^*, σ_2^*) interim utilitarian dominates game form $\tilde{G}$ with the consistent Bayesian equilibrium for all type spaces $(\tilde{\sigma}_1^*, \tilde{\sigma}_2^*)$ if the former interim Pareto dominates the latter.

5.3 Random Dictatorship is Interim Utilitarian Dominated

This paper compares the efficiency of game forms with belief independent equilibria with those of game forms with belief dependent equilibria. (See chapter 4 for a discussion of the motivation for this question.) Consequently we will need to define belief independent equilibria.

Definition V.9. A game form G and a Bayesian equilibrium of G for every type space, (σ_1^*, σ_2^*), is *belief independent* if for all $i \in \{1, 2\}, \mathcal{T}, \widetilde{\mathcal{T}} \in \Upsilon, t_i \in T_i$ and $\tilde{t}_i \in \tilde{T}_i$ such that $u_i(t_i) = \tilde{u}_i(\tilde{t}_i)$ we have:

$$\sigma_i^*(\mathcal{T}, t_i) = \sigma_i^*(\widetilde{\mathcal{T}}, \tilde{t}_i), \quad (5.6)$$

where T_i, u_i correspond to $\mathcal{T}$ and $\tilde{T}_i, \tilde{u}_i$ correspond to $\widetilde{\mathcal{T}}$.

Chapter 4's Proposition IV.9 shows that all game forms and belief independent

equilibria of these game forms that satisfy a pair of unanimity requirements are random dictatorships. To define random dictatorships we need some notation. If u is a utility function, we denote by $b(u)$ the element of A that maximizes u.[3]

We refer to random dictatorship, we shall mean the following specific game form G and specific equilibrium (σ_1^*, σ_2^*) of G for every type space.

Definition V.10. The following game form G and equilibrium (σ_1^*, σ_2^*) of G for every type space will be referred to as *p-random dictatorship*:

(i) $S_1 = S_2 = A$;

(ii)

$$x(s_1, s_2, a) = \begin{cases} 1 & \text{if } s_1 = s_2 = a; \\ p & \text{if } s_1 = a \text{ and } s_2 \neq a; \\ 1-p & \text{if } s_1 \neq a \text{ and } s_2 = a; \\ 0 & \text{if } s_1 \neq a \text{ and } s_2 \neq a; \end{cases}$$

(iii) $\sigma_i^*(\mathcal{T}, t_i, b(u_i(t_i))) = 1$ for all $i \in \{1, 2\}$, $\mathcal{T} \in \Upsilon$, and $t_i \in T_i$.

It is immediate that (σ_1^*, σ_2^*) is a Bayesian equilibrium of G for every type space, and that G and this equilibrium are a random dictatorship. There are other game forms and equilibria that are random dictatorships, but it is without loss of generality to only consider the one described in Definition V.10.

Our first result says that for every $p \in (0, 1)$ such that $p \neq 0$ and $p \neq 1$ there are a game form, and a Bayesian equilibrium of this game form for every type space, that interim Pareto dominate random dictatorship when the probability of agent 1 being dictator is p. We refer to the game form as *p-random dictatorship with compromise.*

[3]Recall that we have assumed that there are no indifferences. Therefore, there is a unique element of A that maximizes u.

Definition V.11. The following game form is called a *p-random dictatorship with compromise.*

(i) for every $i \in \{1, 2\}$:

$$S_i = 2^A \times A,$$

where 2^A is the set of all non-empty subsets of A;

(ii) If $s_1 = (\mathcal{A}_1, a_1)$, $s_2 = (\mathcal{A}_2, a_2)$, and $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$ or $a_1 = a_2$, then:

$$x(s_1, s_2, a) = \begin{cases} 1 & \text{if } a_1 = a_2 = a; \\ p & \text{if } a_1 = a \text{ and } a_2 \neq a; \\ 1 - p & \text{if } a_1 \neq a \text{ and } a_2 = a; \\ 0 & \text{if } a_1 \neq a \text{ and } a_2 \neq a; \end{cases}$$

(iii) If $s_1 = (\mathcal{A}_1, a_1)$, $s_2 = (\mathcal{A}_2, a_2)$, and $\mathcal{A}_1 \cap \mathcal{A}_2 \neq \emptyset$, and $a_1 \neq a_2$, then there is some $a \in \mathcal{A}_1 \cap \mathcal{A}_2$ such that

$$x(s_1, s_2, a) = 1.$$

In words, this game form offers each agent i the opportunity to nominate one preferred alternative, a_i, and also a set $\mathcal{A}_i$ of "acceptable" alternatives. If both agents nominate the same preferred alternative, then it is chosen with probability one. Otherwise, if there is at least one alternative that both voters include in their set of acceptable alternatives, then some alternative that both agents have indicated as acceptable is chosen. If neither of those conditions is met, the mechanism reverts to random dictatorship. We refer to this game form as *random dictatorship with compromise* because it offers agents the opportunity to compromise on a mutually acceptable alternative in place of random dictatorship.

Proposition IV.13 of chapter 4 shows that p-random dictatorship with compromise has a consistent equilibrium for all type spaces (σ_1^*, σ_2^*) that interim Pareto dominates p-random dictatorship and that respects unanimity. This leads to the following proposition.

Proposition V.12. *For all $p \in (0,1)$, p-random dictatorship with compromise has a consistent equilibrium for all type spaces σ^* that interim utilitarian dominates p-random dictatorship and that satisfies positive and negative unanimity.*

Proof. This is an immediate consequence of proposition IIV.13 in chapter 4 and the observation that interim ex post dominance implies interim utilitarian dominance. □

5.4 No Game Form Ex Post Utilitarian Dominates Random Dictatorship

We now show that random dictatorship is ex post utilitarian undominated.

Proposition V.13. *For all $p \in [0,1]$, there is no game form G that has a consistent equilibrium for all type spaces (σ_1^*, σ_2^*) that ex post utilitarian dominates p-random dictatorship.*

Proof. Step 1: We show for every game form G and every equilibrium of G for all type spaces, (σ_1^*, σ_2^*), if G and (σ_1^*, σ_2^*) ex post utilitarian dominate p-random dictatorship, then:

$$\sum_{s_1 \in S_1, s_2 \in S_2} x(s_1, s_2, a) \cdot \sigma_1^*(\mathcal{T}, t_1, s_1) \cdot \sigma_2^*(\mathcal{T}, t_2, s_2) \leq 1 - p \tag{5.7}$$

for all $\mathcal{T} \in \Upsilon$, every $(t_1, t_2) \in T_1 \times T_2$, and every $a \in A$ such that $a \neq b(u_1(t_1))$, and

$$\sum_{s_1 \in S_1, s_2 \in S_2} x(s_1, s_2, a) \cdot \sigma_1^*(\mathcal{T}, t_1, s_1) \cdot \sigma_2^*(\mathcal{T}, t_2, s_2) \leq p \tag{5.8}$$

for all $\mathcal{T} \in \Upsilon$, every $(t_1, t_2) \in T_1 \times T_2$, and every $a \in A$ such that $a \neq b(u_2(t_2))$.

That is, any alternative that is not agent 1's preferred alternative can be chosen with a probability of at most $1-p$, and any alternative that is not agent 2's preferred alternative can be chosen with a probability of at most p. We prove this statement only for agent 1. The proof for agent 2 is analogous.

The proof is indirect. Suppose there were some type space $\mathcal{T}^*$, some $(t_1^*, t_2^*) \in T_1^* \times T_2^*$, and some alternative $a^* \in A$ such that $a^* \neq b(u_1(t_1^*))$, and yet:

$$\sum_{s_1 \in S_1, s_2 \in S_2} x(s_1, s_2, a^*) \cdot \sigma_1^*(\mathcal{T}^*, t_1^*, s_1) \cdot \sigma_2^*(\mathcal{T}^*, t_2^*, s_2) > 1 - p. \tag{5.9}$$

We now construct a new type space, $\widehat{\mathcal{T}}$, and show that in this type space there is a vector of types such that the outcome prescribed by the equilibrium (σ_1^*, σ_2^*) yields lower ex post utilitarian welfare than p-random dictatorship. This contradicts the assumption that G and (σ_1^*, σ_2^*) ex post utilitarian dominate p-random dictatorship.

The type sets in $\widehat{\mathcal{T}}$ are given by: $\widehat{T}_1 = T_1^*$, and $\widehat{T}_2 = T_2^* \cup \{t_2(1), \ldots, t_2(K)\}$ where $K \in \mathbb{N}$ is large enough. We define later how large K needs to be. The types that are contained in T_1^* or T_2^* have the same utility function and beliefs in $\widehat{\mathcal{T}}$ as in $\mathcal{T}$. For types $t_2 \in \{t_2(1), t_2(2), \ldots t_2(K)\}$ the beliefs are given by:

$$\pi_2(t_2(k), t_1^*) = 1. \tag{5.10}$$

The utility function of types $t_2 \in \{t_2(1), t_2(2), \ldots t_2(K)\}$ is:

$$u_2(t_2(k), a) = \begin{cases} 1 & \text{if } a = a^*; \\ \frac{k}{K} & \text{if } a = b(u_1(t_1^*)); \\ 0 & \text{otherwise.} \end{cases} \tag{5.11}$$

This concludes the construction of $\widehat{\mathcal{T}}$.[4] By the consistency of the Bayesian equilibrium

[4]The construction violates our earlier assumption that there are no indifferences. The construc-

(σ_1^*, σ_2^*), for all types in T_1 and T_2, σ_1^* and σ_2^* have to prescribe the same strategies for $\widehat{\mathcal{T}}$ as for $\mathcal{T}^*$. For types $t_2 \in \{t_2(1), t_2(2), \ldots t_2(K)\}$ the strategy $\sigma_2^*(\widehat{\mathcal{T}}, t_2)$ must be a best response to $\sigma_1^*(\widehat{\mathcal{T}}, t_1^*)$.

We denote for every $k \in \{0, 1, 2, \ldots, K\}$ by $v_2(k)$ the expected utility of type $t_2(k)$ in the game form G if equilibrium (σ_1^*, σ_2^*) is played. By standard incentive compatibility arguments $v_2(k)$ is increasing in k. Observe that, for $k \in \{1, 2, \ldots, K\}$, the difference $v_2(k) - v_2(k-1)$ cannot be more than $1/K$ because, by adopting type $t_2(k)$'s strategy, type $t_2(k-1)$ can always get within $1/K$ of type $t_2(k)$'s expected utility. We also denote for $k \in \{0, 1, 2, \ldots, K\}$ by $r_2(k)$ the equilibrium expected utility of type $t_2(k)$ under random dictatorship. It is immediate that $r_2(k)$ is increasing in k, and that $r_2(k) - r_2(k-1) = 1/(pK)$ for $k = 1, 2, \ldots, K$.

Now consider the difference: $v_2(k) - r_2(k)$. The observations of the previous paragraph imply that as k increases the change in the absolute value of this difference, $|(v_2(k) - r_2(k)) - (v_2(k-1) - r_2(k-1))|$, is at most $1/K$. Note that by choosing K large enough, we can make the step size of changes of this difference arbitrarily small. Observe that $v_2(0) > r_2(0)$ because, by the assumption of the indirect proof, in the game form G, type $t_2(k)$ has a strategy that implies that alternative a^* is chosen with a probability larger than $1 - p$, so that in equilibrium type $t_2(0)$ must obtain alternative a^* with at least that probability. By contrast, under random dictatorship, alternative a^* is chosen with probability $1 - p$ only. On the other hand, $v_2(K) \leq r_2(K)$, because random dictatorship yields for agent $t_2(K)$ at least one of his top alternatives with probability 1. What we have said so far implies that we can find some $k \in \{0, 1, 2, ..., K\}$ such that $v_2(k) - r_2(k)$ is strictly positive but arbitrarily close to zero, provided we choose K large enough.

tion and the argument that follows below can easily be modified to comply with this assumption by assigning the bottom ranked alternatives *almost* the same, but not *exactly* the same utility.

Next we note that $v_2(k) > r_2(k)$ implies that

$$\sum_{s_1 \in S_1} x(s_1, s_2, a^*) \cdot \sigma_1^*(\widehat{\mathcal{T}}, t_1^*, s_1) > 1 - p \tag{5.12}$$

for every pure strategy $s_2 \in S_2$ in the support of $\sigma_2^*(\widehat{\mathcal{T}}, t_2(k))$. This is because $v_2(k) > r_2(k)$ implies that every strategy in the support of $\sigma_2^*(\widehat{\mathcal{T}}, t_2(k))$ must yield strictly higher expected utility for type $t_2(k)$ than p-random dictatorship would give to this type. Moreover, the only way in which type $t_2(k)$ can be better off under G and (σ_1^*, σ_2^*) than under p-random dictatorship, where $b(u_1(t_1^*))$ and a^* are chosen with probabilities p and $1-p$ respectively, is by raising the probability of a^* above $1-p$.

Next, we denote for every $k \in \{0, 1, 2, \ldots, K\}$ by $v_1(k)$ the expected utility of type t_1^* when he encounters type $t_2(k)$, and we denote by $r_1(k)$ the expected utility under p-random dictatorship of type t_1^* when he encounters type $t_2(k)$. We first observe that whenever $v_2(k) > r_2(k)$ we must have: $v_1(k) < r_1(k)$. This is because p-random dictatorship would give $b(u_1(t_1^*))$ and a^* with probability p and $1-p$. By contrast, the game form G gives in equilibrium a^* with a probability that is larger than $1-p$. Therefore, the outcome will be worse than random dictatorship for player 1. Now consider all pure strategies of player 2 that, matched with type t_1^*'s equilibrium strategy, yield a probability of a^* of more than $1-p$. As observed before, $v_2(k) > r_2(k)$ implies that type $t_2(k)$ can only play such strategies with positive support. Against each of these strategies player 1 obtains a maximum utility strictly lower than $r_1(k)$. Therefore, there is $\ell > 0$ such that $v_2(k) > r_2(k)$ implies: $v_1(k) < r_1(k) - \ell$.

Now choose K large enough so that we can find a type $t_2(k)$ for whom $v_2(k) > r_2(k)$, but $v_2(k) < r_2(k) + \ell$. We then have: $v_1(k) < r_1(k) - \ell$, and therefore, adding the last two inequalities: $v_1(k) + v_2(k) < r_1(k) + v_r(k)$. This contradicts the hypothesis that G and (σ_1^*, σ_2^*) ex post utilitarian dominate p-random dictatorship.

*S*tep 2: We now complete the proof by showing that no game form G and equilibrium (σ_1^*, σ_2^*) of G for all type spaces that has have the properties described in Step 1 can ex post utilitarian dominate p-random dictatorship. The proof is indirect. Suppose there were some game form G and some equilibrium (σ_1^*, σ_2^*) of G for all type spaces that have the properties described in Step 1 and that ex post utilitarian dominate p-random dictatorship. Then there must be some type space $\mathcal{T}^{**}$ and some $(t_1^{**}, t_2^{**}) \in T_1^{**} \times T_2^{**}$such that:

$$\sum_{i\in\{1,2\}} \sum_{s_1\in S_1, s_2\in S_2} u_i(t_i^{**}, x(s_1,s_2)) \cdot \sigma_1^*(\mathcal{T}^{**}, t_1^{**}, s_1) \cdot \sigma_2^*(\mathcal{T}^{**}, t_2^{**}, s_2) > pu_1(b(u_1(t_1^{**}))) + (1-p)u_2(b(u_2(t_2^{**}))) \tag{5.13}$$

We now construct a new type space, $\widetilde{\mathcal{T}}$, and show that in this type space there is a vector of types such the outcome prescribed by the equilibrium (σ_1^*, σ_2^*) yields lower utilitarian welfare than p-random dictatorship. This contradicts the assumption that G and (σ_1^*, σ_2^*) ex post utilitarian dominate p-random dictatorship. The construction and the argument below are very similar to, but not identical to, the argument in Step 1.

Before we begin the construction we note that it must be that in equilibrium, at t^{**}, either $b(u_1(t^{**}))$ is chosen with probability strictly less than p, or $b(u_2(t^{**}))$ is chosen with probability strictly less than $1-p$, or both. Otherwise, the game form G with the equilibrium (σ_1^*, σ_2^*) could not yield strictly higher utilitarian welfare at t^{**} than p-random dictatorship. Without loss of generality, we focus on the case that $b(u_1(t^{**}))$ is chosen with probability strictly less than p. The other case can be dealt with by a symmetric argument. Let a^{**} be the second most preferred alternative of agent 1 at t_1^{**}.

We now construct $\widetilde{\mathcal{T}}$. The type sets are given by: $\widetilde{T}_1 = T_1^{**}$, and $\widetilde{T}_2 = T_2^{**} \cup \{t_2(1), \ldots, t_2(K)\}$ where $K \in \mathbb{N}$ is large enough. We define later how large K needs

to be. The types that are contained in T_1^{**} or T_2^{**} have the same utility functions and beliefs in $\widetilde{\mathcal{T}}$ as in $\mathcal{T}^{**}$. For types $t_2 \in \{t_2(1), t_2(2), \ldots t_2(K)\}$ the beliefs are given by:

$$\pi_2(t_2(k), t_1^{**}) = 1. \tag{5.14}$$

The utility function of types $t_2 \in \{t_2(1), t_2(2), \ldots t_2(K)\}$ is:

$$u_2^*(t_2(k), a) = \begin{cases} 1 & \text{if } a = a^{**}; \\ \frac{k}{K} & \text{if } a \neq b(u_1(t_1^{**})) \text{ and } a \neq a^{**}; \\ 0 & \text{if } a = b(u_1(t_1^{**})). \end{cases} \tag{5.15}$$

This concludes the construction of $\widetilde{\mathcal{T}}$.[5] By the consistency of the Bayesian equilibrium (σ_1^*, σ_2^*), for all types in T_1^{**} and T_2^{**}, σ_1^* and σ_2^* have to prescribe the same strategies for $\widetilde{\mathcal{T}}$ as for $\mathcal{T}^{**}$. For types $t_2 \in \{t_2(1), t_2(2), \ldots t_2(K)\}$ the strategy $\sigma_2^*(\widetilde{\mathcal{T}}, t_2)$ must be a best response to $\sigma_1^*(\widetilde{\mathcal{T}}, t_1^{**})$.

We denote for every $k \in \{0, 1, 2, \ldots, K\}$ by $v_2(k)$ the equilibrium expected utility of type $t_2(k)$ in the game form G with equilibrium (σ_1^*, σ_2^*). By standard incentive compatibility arguments $v_2(k)$ is increasing in k. Observe that, for $k \in \{1, 2, \ldots, K\}$, the difference $v_2(k) - v_2(k-1)$ cannot be more than $1/K$ because, by adopting type $t_2(k)$'s strategy, type $t_2(k-1)$ can always get within $1/K$ of type $t_2(k)$'s expected utility. We also denote for $k \in \{0, 1, 2, \ldots, K\}$ by $r_2(k)$ the equilibrium expected utility of type $t_2(k)$ under random dictatorship. It is immediate that $r_2(k) = 1 - p$ for all $k = 1, 2, \ldots, K$.

Now consider the difference: $v_2(k) - r_2(k)$. The observations of the previous paragraph imply that as k increases the difference increases, and that moreover it can change by at most $1/K$. Note that by choosing K large enough, we can make the step

[5]The construction violates our earlier assumption that there are no indifferences. The construction and the argument that follows below can easily be modified to comply with this assumption by assigning to the middle ranked alternatives *almost* the same, but not *exactly* the same utility.

size of changes of this difference arbitrarily small. Observe next that $v_2(0) \leq r_2(0)$. This is because under G and (σ_1^*, σ_2^*) alternative a^{**} can be chosen with a probability of at most $1-p$, by Step 1 of this proof applied to player 1. Therefore, $v_2(0) \leq 1-p = r_2(0)$. Finally, we show that $v_2(K) > r_2(K)$. Observe that $v_2(K) > 1-p$, because, by assumption, the probability of $b(u_1(t_1^{**}))$ under G and (σ_1^*, σ_2^*) at (t_1^{**}, t_2^{**}) is strictly less than p. Thus the probability of all other alternatives together must be strictly more than $1-p$. Type $t_2(K)$ can choose the same strategy as type t_2^{**}, and therefore, if type $t_2(K)$ chooses optimally, $v_2(K) > 1-p = r_2(1)$. What we have said so far implies that we can find some $k \in \{0, 1, 2, ..., K\}$ such that $v_2(k) - r_2(k)$ is strictly positive but arbitrarily close to zero, provided we choose K large enough.

Next we note that $v_2(k) > r_2(k)$ implies that

$$\sum_{s_1 \in S_1} x(s_1, s_2, b(u_1(t^{**}))) \cdot \sigma_1^*(\widetilde{\mathcal{T}}, t_1^{**}, s_1) < p \tag{5.16}$$

for every pure strategy $s_2 \in S_2$ in the support of $\sigma_2^*(\widetilde{\mathcal{T}}, t_2(k))$. This is because every strategy in the support of player 2's strategy $\sigma_2^*(\widetilde{\mathcal{T}}, t_2(k))$ must yield the same expected utility, and hence strictly higher expected utility than $r_2(k)$. But if such a strategy implements $b(u_1(t^{**}))$ with probability of p or more, then the remaining probability that is distributed among a^{**} and all other alternatives, is at most $1-p$. Therefore, player 2's expected utility from such a strategy is no more than $1-p = r_2(k)$, which contradicts our assumption that player 2's expected utility is more than $r_2(k)$.

Next, we denote for every $k \in \{0, 1, 2, \ldots, K\}$ by $v_1(k)$ the expected utility of type t_1^{**} when he encounters type $t_2(k)$, and we denote by $r_1(k)$ the expected utility under random dictatorship of type t_1^{**} when he encounters type $t_2(k)$. We first observe that whenever $v_2(k) > r_2(k)$ we must have: $v_1(k) < r_1(k)$. This is because random dictatorship would give $b(u_1(t_1^*))$ and a^{**} with probability p and $1-p$. By contrast, the

game form G gives in equilibrium $b(u_1(t_1^{**}))$ with a probability that is less p. Therefore, the outcome will be worse than random dictatorship for player 1. Now consider all pure strategies of player 2 that, matched with type t_1^*'s equilibrium strategy, yield a probability of $b(u_1(t_1^{**}))$ of strictly less than p. As observed before, $v_2(k) > r_2(k)$ implies that type $t_2(k)$ can only play such strategies with positive support. Against each of these strategies player 1 obtains a maximum utility strictly lower than $r_1(k)$. Therefore, there is $\ell > 0$ such that $v_2(k) > r_2(k)$ implies: $v_1(k) < r_1(k) + \ell$.

Now choose K large enough so that we can find a type $t_2(k)$ for whom $v_2(k) > r_2(k)$, but $v_2(k) < r_2(k) + \ell$. We then have: $v_1(k) < r_1(k) - \ell$, and therefore, adding the last two inequalities: $v_1(k)+v_2(k) < r_1(k)+v_r(k)$. This contradicts the hypothesis that G and (σ_1^*, σ_2^*) ex post utilitarian dominate p-random dictatorship. □

5.5 Conclusion

This paper extends the results of chapter 4 to the case of a mechanism designer whose preferences obey relative utilitarianism and who faces two agents. For a mechanism designer who is concerned with relative utilitarianism and considers interim expected utilities, the same mechanism that is used in that paper, random dictatorship with compromise, is more efficient than random dictatorship. This result is an immediate implication of the interim Pareto result in that paper and the definition of interim relative utilitarianism, and consequently extends to the more than two agent case. The other main result is that no mechanism dominates random dictatorship from an ex post relative utilitarian perspective. It is not immediately clear how to extend the argument to more than two agents. Consequently, whether any mechanism dominates random dictatorship from an ex post relative utilitarian perspective when there are more than two agents remains an open question.

References

Amrita Dhillon (1998), Extended Pareto Rules and Relative Utilitarianism, *Social Choice and Welfare* 15, 521-542.

Amrita Dhillon and Jean Francois Mertens (1999), Relative Utilitarianism, *Econometrica* 67, 471-498.

CHAPTER VI

The Role of Solidarity and Reputation Building in Coordinating Collective Resistance

6.1 Introduction

What motivates agents to resist when a leader attempts to gain or maintain influence via a divide-and-conquer strategy? Examples of leaders breaking up larger groups into smaller ones that have less power are prevalent — from the times of Machievelli (Zeitlin and Weyher, 2001), to Western countries using the strategy in Africa (Croucher, 2004), to Wal-mart's recent pledge to "go green" (Tasini, 2008), to management/labor disputes (LeDuff, 2000). Often the strategy is successful, but it also is often met by *coordinated joint resistance* — subordinates show solidarity, banding together at a personal cost to thwart the divide-and-conquer strategy. Examples include strikes and unionization (Horowitz, 1997; Zeitlin and Weyher, 2001; Gordon and Lenhardt, 2007; Oyogoa, 2009). Our research uses experimental methods to examine the extent to which this type of subordinate solidarity is driven by a fairness norm or driven by the desire to build expectations among both leaders and other subordinates that leader transgressions will be successfully resisted.

Weingast (1997) introduced a political model representing how exploitative leaders can effectively maintain power in a society where there are different interest groups. It

also has natural applications to other situations where one individual has authority over multiple subordinates; for example, a manager and worker relationship. The situation can be modeled by the "coordinated resistance" (CR) game.[1] In the CR game, there are three players: one leader and two subordinates, A and B. The leader moves first and has four available actions: (i) transgress against both subordinates, (ii) transgress against neither, (iii) transgress against A, and (iv) transgress against B. These last two actions provide the leader an opportunity to "divide-and-conquer". The subordinates observe the leader's action and then have two available actions: challenge or acquiesce. As long as both subordinates do not challenge, a "divide-and-conquer" strategy gains rewards for the leader at the expense of the targeted subordinate. The beneficiary of the transgression (i.e. the subordinate not targeted) earns a higher payoff by acquiescing than by challenging, regardless of what action the targeted subordinate takes. For a small cost, though, the beneficiary can challenge the leader's transgression, and if the targeted subordinate also challenges they achieve *coordinated joint resistance*. At this outcome both subordinates earn what they would get if the leader did not transgress minus the small cost of challenging, and the leader earns zero, which is significantly less than if he had chosen transgress against neither. Social surplus is maximized when the leader chooses transgress against neither and both subordinates acquiesce. However, this outcome is not part of any equilibrium.

Cason and Mui (2007) examine the CR game in a laboratory experiment, focusing on when communication does or does not allow subordinates to successfully work together to resist exploitation by a leader. They use random anonymous rematching between repeated plays, and find that when the leader chooses to transgress against one subordinate, a significant fraction of the subordinates coordinate on joint resistance. This is despite the beneficiary paying a price in their own period payoff to help the targeted subordinate with no direct material benefit. Over time, some leaders

[1] Weingast called it the" Sovereign-Constituency Transgression Game".

adapt their strategy, choosing to transgress against neither. Cason and Mui interpret beneficiaries challenging, and hence coordinated joint resistance, as evidence of fairness: some subordinates are "altruistic punishers" (Fehr and Gächter, 2002; Boyd et al, 2003; DeQuervain et al, 2004; Fehr and Fischbacher, 2004; Gintis et al, 2005). Subordinates choose to punish the leader — even at a personal cost — for having violated social norms of fairness.[2]

An alternative argument — not invoking an appeal to social preferences — is that subordinates have an individual incentive to create an expectation among leaders that "divide-and-conquer" style transgression will be met with coordinated joint resistance. After all, when individuals are randomly and anonymously rematched each subordinate has an equal chance of being the victim of transgression in the future. If there is a greater than $\frac{1}{4}$ probability of joint resistance, risk-neutral leaders will not want to transgress. With a low cost of resisting, subordinates may regard investing in a group reputation for joint resistance as a worthwhile strategy. If the subordinates can get the leader to choose the transgress against neither option, they can secure a higher payoff than when they face a 50/50 chance of being the victim of the leader's transgression. This suggests a very different motive for joint resistance: a subordinate has a strategic motive to build a reputation as one who challenges in order to alter the leader's or the other subordinate's behavior in the future.

We use a novel experimental design to test the extent to which repeated interactions drive some subordinates to resist when they are the current beneficiaries of the leader's divide-and-conquer strategy. We do this by systematically reducing the strategic incentive to resist. The first treatment has leaders pre-commit to their behavior in all periods. The actions the leaders commit to cannot be conditioned on any play by the subordinates. Such leader commitment removes the ability for sub-

[2]Evidence that individuals possess a taste for punishment has been provided across a variety of experimental games, including public goods (Fehr and Gächter, 2000; Fehr and Gächter, 2002), investment games (DeQuervain et al, 2004, Rigdon 2009), and ultimatum games (Güth et al, 1985; Forsythe et al, 1994, Xiao and Houser, 2005).

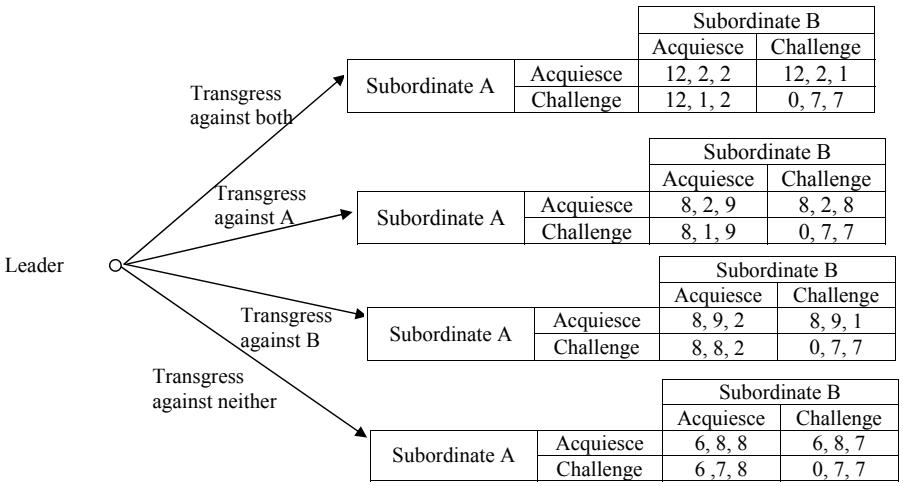


Figure 6.1: CR Game (payoffs are Leader, Subordinate A, and Subordinate B)

ordinates to influence the leaders' future behavior by their action today. The second treatment goes one step further: leaders pre-commit to their behavior in all periods, as in the first treatment, and one subordinate also pre-commits to their behavior in all periods. The pre-commited subordinate's commited actions cannot be conditional on other agent's actions in other rounds. Such subordinate commitment removes the ability of subordinates to alter behavior of the leader or the behavior of the subordinate with whom they interact. Each of these treatments leave fairness considerations intact. We can then compare the rates of beneficiary challenge and coordinated joint resistance across the treatments. By comparing these treatments with a baseline treatment where both motivations may be at work, we aim to disentangle the effects of other-regarding motives from those of reputation-building and coordination incentives.

The next section describes the features of the CR game and experimental findings to date. Section 6.5 details the experimental design and outlines the empirical hypotheses. Section 6.6 describes the procedures in the experiment and Section 6.7 discusses the results. The final section concludes.

6.2 The Coordinated Resistance Game

Our experimental design uses the version of the leader-subordinate scenario drawn from prior experimental literature (Cason and Mui, 2007,Cason and Mui, 2009); see Figure 6.1. Of the leader's four actions, two treat the subordinates symmetrically (Transgress Against Both, Transgress Against Neither) and two treat the subordinates asymmetrically (Transgress Against A, Transgress Against B). If the leader chooses Transgress Against Both — rather than Transgress Against Neither — and the subordinates do not both challenge, the leader's payoff is increased by 6 and both subordinates' payoffs are lowered by 6 units. In the case of Transgress Against A, the leader gains 2 units and the subordinate transgressed against (A) has her payoff reduced by 6 units, exactly as in the Transgress Against Both case; A is referred to as the *targeted* subordinate. The subordinate not transgressed against (B) has her payoff raised by 1 unit relative to the Transgress Against Neither case; this represents a bribe by the leader and so B is referred to as the *beneficiary*. The case of Transgress Against B works identically, except B is targeted and A is the beneficiary. A beneficiary's dominant strategy is to choose Acquiesce. However, if the transgression by the leader is to be resisted, both subordinates must choose challenge — *coordinated resistance* corresponds to the lower-right square in each matrix (payoffs: $0, 7, 7$).

There are three pure strategy subgame perfect Nash equilibria in the CR game: (i) the leader plays Transgress Against Both, and both subordinates play Acquiesce; (ii) the leader plays Transgress Against A and both subordinates play Acquiesce; and (iii) the leader plays Transgress Against B and both subordinates play Acquiesce. Importantly, there is no equilibrium of the game that involves the leader playing Transgress Against Neither, and there is no equilibrium that has the subordinates coordinating on the action Challenge when the leader attempts to divide and conquer the subordinates.

6.3 Prior Experimental Results

Cason and Mui (2007) have subjects play the CR game with the same payoffs we use, and examine the effects of communication on the level of coordinated resistance achieved by subordinates. In the *baseline*, subordinates do not communicate with the leader or each other prior to making their decisions. There is a low frequency of both coordinated resistance and leaders choosing Transgress Against Neither. In the *ex-post communication* treatment, subordinates observe the leader's action and then send signals to the other subordinate indicating what action they intend to take given the leader's action. The likelihood of coordinated resistance was higher than with no communication, with beneficiaries challenging roughly half of the time when both subordinates signaled an intention to challenge. Additionally, leaders played Transgress Against Neither significantly more often in the last thirty periods, roughly 25%. In the *private ex-ante* communication treatment, prior to observing the leader's choice, subordinates send signals to each other indicating what choice they intend to make for each of the four leader's actions; subordinates then learn of the leader's action, and the subordinates simultaneously decide on an action which could be the same or different from their intended choice. Beneficiaries challenged about 33.2% of the time when both indicated an intent to challenge, and leaders played Transgress Against Neither at a significantly higher rate compared to any other treatment, roughly 37%.

Cason and Mui (2009) extend the experimental design, examining how repeated matching with various ending rules interacts with the form of communication allowed between subordinates to impact rates of coordinated resistance and leader transgression. Overall, they find that communication is better than repetition in coordinating resistance. They suggest this is because "it makes it easier for subordinates to identify others who have social preferences and are willing to incur the cost to punish a violation of social norms (p. 1)". Next, we describe such a hypothesis more clearly

and put forward an alternative hypothesis that explains the current data equally well.

6.4 Why Do Beneficiaries Challenge Transgression?

One hypothesis for the high rate of beneficiaries challenging is that some subordinates have a preference for fairness. They care about the outcomes of the other players, and hence are willing to challenge when they are a beneficiary because they prefer the outcome in that case (lower payoff for leader, higher payoff for victim, slightly lower payoff for themselves) to the outcome when they acquiesce (higher payoff for leader, lower payoff for victim, slightly higher payoff for themselves). Hence, they are willing to pay an economic cost to challenge the leader's transgression against the victim. Such subordinates have a social preference reflecting solidarity with other subordinates, and are willing to punish a leader viewed as violating a social norm when he attempts to divide-and-conquer. This form of punishment can aid in the maintenance of norms (Ostrom, 1990; Fehr and Gächter, 2002; Fehr and Fischbacher, 2003; Egas and Riedl, 2008).

This paper introduces another hypothesis: the high rate of beneficiaries challenging — and high rate of coordinated joint resistance — is due to strategic decision-making by the subordinates. There are at least two strategic reasons for a beneficiary to challenge transgression. The first reason is vertical in nature, and the hypothesis is that the beneficiary choosing challenge aims to change the expectations of a leader and hence alter his decision to one more beneficial in the future. The second reason is horizontal in nature, and the hypothesis is that the beneficiary choosing challenge aims to change the expectations of the other subordinate and hence alter her behavior in the future. In both cases, there is strategic incentive to create an expectation that transgression will be met with resistance. This relies on the argument that the observations of anonymous individuals' play can affect beliefs about the *distribution of types* in the population, and hence change future behavior through

updating. The first reason — vertical in nature — is that if the leader's divide-and-conquer strategy gets met with resistance, then the leader may change his strategy in the future to one where he chooses the Transgress Against Neither option. Under this option, subordinates can secure a higher payoff than when they face a 50/50 chance of being the victim of the leader's transgression. Subordinates receive a certain payoff of 8 units under Transgress Against Neither compared to an expected payoff of $\frac{1}{2}(9) + \frac{1}{2}(2) = 5.50$ units from Transgress Against A or Transgress Against B. The second reason — horizontal in nature — is that by challenging, a subordinate can signal to the other subordinate that resistance can be successful in an effort to alter the other subordinates behavior in the future.[3] Observing a beneficiary challenging is a very strong signal, and can influence another subordinate's strategy from one of acquiescing to one of challenging. If so, then the pair can reach coordinated joint resistance more often in the future and thereby earn higher payoffs. Both kinds of strategic decision-making could be at work in the standard CR game. As a result, it is unclear the extent to which coordinated resistance is motivated by expectation building and the extent to which it is motivated by social preferences of subordinates. Our experiment systematically removes each of the strategic reasons for a beneficiary to challenge transgression while leaving in tact the motives for solidarity. Each treatment involves some portion of the subjects pre-committing their actions. We explain this in more detail in the next section.

6.5 Experimental Design and Hypotheses

The baseline (B) treatment is a replication of the *private ex-ante* communication treatment implemented by Cason and Mui (2007). One treatment — Leader Com-

[3]The expected future benefit to subordinates will be smaller under random rematching than under repeated interactions since it may be many periods before interacting with the same leader or same subordinate again, and with complete anonymity the subordinate will not know when they do.

mitment (LC) — has leaders pre-commit to their decisions in all periods. Such leader commitment removes the ability for subordinates to influence the leaders' future behavior by their action today. The second treatment — Subordinate Commitment (SC) — goes one step further: leaders pre-commit to their decisions in all periods as in LC and one subordinate also pre-commits to their intended actions and decisions in all periods. Subordinate commitment removes the ability of the freely choosing subordinate to alter behavior of the leader *and* the ability to alter the behavior of the subordinate with whom they interact. Each of these treatments leave fairness considerations intact.

The interaction between the leader and the two subordinates is as follows. The leader chooses an action. Prior to learning the leader's decision, subordinates are able to communicate a message indicating their intended choice of challenge or acquiesce for each of the four possible actions by the leader. Each subordinate then observes the four intended choices signaled by the other subordinate. Then, the subordinates learn the leader's decision, and simultaneously choose an action to challenge or acquiesce; the subordinate's chosen action can be the same or different from the intended choice specified under that action. The leader and subordinates learn all of the decisions and the respective payoffs. This concludes the interaction. Each experimental session consists of two phases, Phase 1 and Phase 2.[4] Phase 1 is identical across all of the treatments — subjects interact as described above for 10 periods and are randomly re-matched at the end of each period. Phase 2 consists of the treatment portion of the experiment.

In B, Phase 2 is identical to Phase 1 except subjects participate in 40 periods in the phase, being randomly re-matched at the end of each period. The choice of

[4]The two phase structure of all of our treatments is slightly different from Cason and Mui, where they have subjects participate in one phase consisting of 50 periods total. We implemented Phase 1 so that subjects would have experience interacting in the CR game before having to commit to strategies in the future. It is a variation that we do not expect to make a difference. The instructions and protocol are otherwise identical in the baseline.

this treatment as our baseline is based on two considerations. First, it is Cason and Mui's treatment with the highest level of coordinated resistance by subordinates (making it easier to distinguish whether resistance rates are lower in other treatments compared to the baseline treatment), and also the highest fraction of leaders who choose the Transgress Against Neither action in later periods. Since we are interested in comparing the rates of coordinated resistance across treatments, it makes sense to select a baseline with the highest rate. Second, there is a similarity in the decision-making environments for private ex-ante signals and our treatment SC where one subordinate will pre-commit to signals and conditional actions in future periods before any others' actions are observed. In both cases signals are chosen ex-ante, before the leader's action can be observed.

In LC, the baseline treatment is modified to remove the vertical nature of the strategic incentives from repeated play. This is accomplished by having the leaders pre-commit to a strategy for the duration of the experiment. Subjects begin with Phase 1. In Phase 2, leaders choose *all their actions* for the 40 periods of the phase before any further periods are played. In each period, both subordinates observe the leader's pre-committed strategy for that period (only) and respond. Since the leaders are pre-committed to a course of action regardless of subordinates' choices, this removes the incentive subordinates may have to produce a expectation for resistance among the leaders. Consequently, under the hypothesis that subordinates are challenging to influence the leader's behavior in the future, we predict that beneficiary resistance — and hence coordinated resistance — will be lower in LC than in B. Subordinates will perceive no reputation-building motivation if the leaders are already committed to their future course of action, even though they will play further rounds of the game. Therefore, fewer beneficiaries will choose challenge, and this is our first hypothesis.

Hypothesis 1 (Leader Commitment)**.** Beneficiary resistance will be lower in LC than

in B.

The LC treatment leaves open the possibility that subordinates' actions will still reflect their repeated interaction *with other subordinates*; that is, it leaves open the horizontal nature of strategic decision-making. There is still an incentive to coordinate on resistance during any one period where only one subordinate is targeted because the beneficiary of the transgression may find herself targeted by the leader in a future period. In this sense, there is room for subordinates to influence the expectation of other subordinates. To disentangle this expectation building incentive from other-regarding preferences, we conduct an additional treatment. The SC treatment is identical to LC with the additional constraint that half of the subordinates are also pre-committed to a strategy in Phase 2. One subordinate chooses all of her intended actions and actual actions for the remaining periods before observing any actions by the others. The matching is such that each group contains one pre-committed leader, one pre-committed subordinate, and one subordinate who is free to choose her action period-by-period. The subordinate commits to an action in response to each possible leader action in the period. In addition, the committed response to a specific leader action can be contingent on the other subordinate's intended action in response to that action by the leader in that period (but not other periods; the committed action in the period can only be contingent on that period's leader action and other subordinate's intended action given the leader's action). To simplify the set of choices available to pre-committing subordinates, they are restricted to acquiesce if the leader plays Transgress Against Both. The "free" subordinate knows that they will never interact with a leader or a subordinate whose behavior is not pre-committed. This removes all strategic incentives for the "free" subordinate, but still leaves open the possibility of beneficiary challenging due to other-regarding preferences. Given that the "free" subordinate interacts with players who have already committed to their future course of action, she will perceive no incentive to change the leader's or the subordinate's

expectations as this cannot change either's behavior in the future, even though they will play further rounds of the game. Therefore, we expect that beneficiary resistance will be even lower in this treatment than the others, and this is our second hypothesis.

Hypothesis 2 (Subordinate Commitment)**.** Beneficiary resistance by the "free" subordinates will be lower in SC than by the "free" subordinates in LC.

These are our two primary hypotheses. Define $R(T)$ as the rate of beneficiary resistance under treatment T. Then H_1 and H_2 imply the following predicted ranking:

$$R\,(\mathrm{B}) > R\,(\mathrm{LC}) \geq R\,(\mathrm{SC})$$

These rankings arise from the general hypothesis that by systematically removing strategic incentives in repeated play, a subordinate's willingness to engage in beneficiary resistance will be reduced. Since beneficiary resistance is necessary for subordinates to obtain joint resistance, we also hypothesize that rates of joint coordinated resistance will follow the same pattern across treatments.

6.6 Procedures

The experimental sessions were conducted at the Robert Zajonc's Laboratory in the Institute for Social Research at the University of Michigan.[5] We completed a total of 24 independent sessions with 8 sessions for each of the three experimental conditions. The experiment was computerized using z-tree (Fischbacher, 2007). Participants were undergraduates and were recruited using standard experimental procedures. They participated in only one session. Each session required 1.5 to 2.5 hours to complete.

An experimental session ran as follows. Subjects received a $10 show-up payment and were seated in the laboratory. Two sessions were conducted simultaneously with

[5]We ran the first session in late November 2007 and completed the last session in late April 2009. There were some coding issues that extended the length of time necessary to complete the treatments.

each session having 3 leaders and 6 subordinates.[6] Once all subjects had completed a consent form, the instructions were read aloud. The instructions used were the same as those used by Cason and Mui (2007). The instructions used neutral terms for the roles of leader and subordinates as well as their available actions: "Person 1" (the leader) chooses from "earnings squares" A, B, C or D and "Persons 2 and 3" (the subordinates) simultaneously choose X or Y. Subjects in each treatment received the same instructions for Phase 1 of the session, being told that Phase 1 would last 10 periods (see Appendix **??** for the instructions). Subjects were informed that there would be a second phase to the experiment as follows: "Phase 2 will be a similar decision-making task, and you will have the same role in it that you have in Phase 1. Further instructions will be provided before Phase 2 begins."[7] Subjects were required to complete a quiz which included payoff calculations for all roles and questions to check their understanding about the interaction; the experimenter checked the quizzes for accuracy. The quiz used was the same as that used by Cason and Mui (2007).[8] Subjects were then randomly assigned a role as either a leader or a subordinate, and they kept this role throughout the session. Subjects were randomly re-matched each period. Subjects first participated in Phase 1, consisting of 10 periods of the baseline version of the CR game. Then, prior to beginning Phase 2, further instructions were given: in B subjects were informed they would participate in 40 more periods, identical to those in Phase 1 (see Appendix **??**); in LC the instructions provided additional information that the leader would pre-commit to a strategy for all 40 periods before any more periods were played (see Appendix **??**); and in

[6]Subjects were not aware that the re-groupings only happened among the 9 subjects in their particular session. This is in keeping with Cason and Mui's protocol.

[7]This deviates from Cason and Mui's procedures, which had only one phase of 50 total periods. Our aim is to give subjects experience in the CR game — with decisions counting for monetary payment — prior to selecting pre-committed actions.

[8]Subjects were asked true/false questions about the interaction; for example, "You remain grouped with the same two other participants in all decision-making periods" and "If you are Person 2 or Person 3, you must make the same choice on your decision screen as you indicated in the relevant intention screen" and "If you are Person 2 or Person 3, your intentions are shown to all three people in your group (Person 1, Person 2 and Person 3) before anyone makes actual decisions."

SC the instructions provided additional information that in each three-person group, both the leader and one of the subordinates would be pre-committed while the other subordinate would make decisions as in Phase 1 (see Appendix **??**). Subjects then participated in Phase 2, completing an additional 40 periods of one treatment, being randomly re-matched each period. Participants earned points that were exchanged for dollars at the exchange rate of \$0.09/point. Once the session finished, subjects were paid their accumulated earnings in private. Average earnings (excluding the show-up payment) in B were \$29.76 for the leaders and \$27.51 for subordinates; in LC earnings were \$30.26 for the leaders and \$27.01 for subordinates; and in SC earnings were \$25.05 for the leaders and \$30.36 for subordinates.

6.7 Results

In this section, we begin by analyzing the extent to which the baseline data is similar to the *private ex-ante* treatment conducted by Cason and Mui (2007). As we demonstrate, the results are statistically indistinguishable. We then turn to analyzing the rates of beneficiary and coordinated resistance.

Our focus is on Phase 2 behavior. Recall that Phase 1 is identical across treatments — subjects participate in the baseline version of the CR for 10 periods. We are interested in differences across treatments so the discussion will center on the results in Phase 2. Periods of Phase 2 are labeled 11–50. We further restrict analysis to periods 21–50 to allow time for subjects to learn — in both treatment conditions, the decision environment has changed significantly from the baseline version of Phase 1. To avoid issues of within-session dependence between participants' actions, all the tests we report use each session as an independent observation, so there are eight observations per treatment.

6.7.1 Replication

The baseline treatment uses instructions and procedures nearly identical to those in Cason and Mui's *private ex-ante* treatment. We compare the results along a number of important behavioral dimensions, and find that the two are statistically indistinguishable with respect to the following: the rate at which beneficiaries resist when both indicate an intention to challenge, (27.1%, $p = 0.6818$, two-sided Mann-Whitney test); the rate of joint resistance when both indicate an intention to challenge (24.6%, $p = 0.7718$, two-sided Mann-Whitney test); the rate at which beneficiaries resist given any signal (12.7%, $p = 0.9045$, two-sided Mann-Whitney test); the rate of joint resistance given any signal (10.1%, $p = 0.9522$, two-sided Mann-Whitney test); and the rate of leader non-transgression (24.0%, $p = 0.5619$, two-sided Mann-Whitney test). As a result, we conclude that behavior in the baseline is indistinguishable from that reported by Cason and Mui (2007). We now turn to discussing differences across treatments and testing our hypotheses.

6.7.2 Beneficiary Resistance Rates Across Treatments

This section looks at the frequency of beneficiary resistance and coordinated resistance when the leader attempts to divide-and-conquer, irrespective of the signals sent by the beneficiary and victim. Table 6.2 reports the frequencies for each treatment.

We first briefly look at the correlation between signals and challenging. The data is summarized in Table 6.1. In the baseline treatment, the beneficiary indicated challenge 48.5% of the time. This message was not particularly truthful: when a beneficiary indicated they would challenge, and conditional on the victim also indicating challenge, the beneficiary actually challenged only 27.1% of the time. However, the frequency of beneficiary challenging when both indicated an intention to challenge is significantly higher than under any signal — 27.1% versus 12.7% ($p = 0.00830$, one-sided pairwise t-test). In LC, beneficiaries were somewhat more truthful in their

	Baseline			Leader Commit		
Message: Challenge	Victim Challenges	Beneficiary Challenges	Joint Resistance	Victim Challenges	Beneficiary Challenges	Joint Resistance
Only Victim	52/185 28.1	5/185 2.7	2/185 1.1	60/238 25.2	2/238 0.8	1/238 0.4
Only Beneficiary	24/57 42.1	6/57 10.5	2/57 3.5	22/52 42.3	12/52 23.1	7/52 13.5
Both	145/203 71.4	55/203 27.1	50/203 24.6	115/161 71.4	71/161 44.1	58/161 36.0
Neither	16/91 17.6	2/91 2.2	0/91 0.0	13/115 11.3	1/115 0.9	0/115 0.0

	Subordinate Commit (Free Sub Targeted)			Subordinate Commit (Commited Sub Targeted)		
Message: Challenge	Victim Challenges	Beneficiary Challenges	Joint Resistance	Victim Challenges	Beneficiary Challenges	Joint Resistance
Only Victim	35/69 50.7	15/69 21.7	12/69 17.4	6/28 21.4	1/28 3.6	0/28 0.0
Only Beneficiary	1/3 33.3	0/3 0.0	0/3 0.0	13/39 33.33	11/39 28.2	4/39 10.3
Both	40/41 97.6	26/41 63.4	26/41 63.4	47/49 95.9	9/49 18.4	9/49 18.4
Neither	1/38 2.6	0/38 0.0	0/38 0.0	0/36 0.0	0/36 0.0	0/36 0.0

Table 6.1: Action by Subordinates Conditional on Message of Challenge (Periods 21–50)

signals: they indicated they would challenge 37.6% of the time, and when they indicated challenge they actually challenged (conditional on the victim indicating challenge) 44.1% of the time. This is significantly higher than the rate across any signals, 15.2% ($p = 0.01039$, one-sided pairwise t-test).

In the subordinate commit treatment, when the committed subordinate was targeted the free subordinate indicated they would challenge over half (57.9%) the time but followed through, when the committed subordinate had indicated they would challenge, only 18.4% of the time. This frequency was only marginally significantly higher than for any signals (13.8%, $p = 0.09013$, one-sided pairwise t-test). When the free subordinate was targeted, the committed subordinate only indicated they would challenge 29.1% of the time, but conditional on both indicating challenge they challenged 63.4%, significantly higher than the overall rate of challenge (27.2%, $p = 0.06160$, one-sided pairwise t-test).

As can be seen from Table 6.1, in all treatments victims indicated challenge more

than half the time, and conditional on both subordinates indicating challenge they challenged more than 70% of the time.

Now we move on to our main results. Using the measures of resistance conditional on any signal to test our hypotheses, we find that beneficiary resistance rates in LC are not significantly different from B — 15.2% versus 12.7% ($p = 0.3974$, one-sided Mann-Whitney test).

Result 1. The rate of beneficiary resistance in LC is not significantly different from the rate in B.

The rates are also not different in SC for actions taken by the free subordinate — 13.8% versus 12.7% ($p = 0.2005$, one-sided Mann-Whitney test).

Result 2. The rate of beneficiary resistance in SC when the beneficiary is the free subordinate is not significantly different from the rate in B.

Results 1 and 2 do not support our hypotheses 1 and 2. We also look at the behavior of committed subordinates in SC when they are beneficiaries. The rate of beneficiary challenge for actions taken by the committed subordinate when the free subordinate is the one targeted in SC is over twice the rate of beneficiary challenge in B — 27.2% versus 12.7% — but this difference is due to a couple of extreme sessions, and the difference is not significant when tested by a rank-order test ($p = 0.2483$, one-sided Mann-Whitney test).

Levels of joint resistance (when the leader chooses target transgression) follow the same pattern as beneficiary resistance - the rate of joint resistance in B, 10.7%, is not significantly different from the rate in LC, 11.7% (p = 0.4364, one-sided Mann-Whitney test), and in SC when the beneficiary is the free subordinate - 8.6% (p = 0.2148, one-sided Mann-Whitney test). The level is higher in magnitude, 25.2%, in SC when the beneficiary is the committed subordinate, but the difference is due to a

	Baseline		Leader Commit		Sub Commit Free Targeted		Sub Commit Committed Targeted	
	Ben Chall	Joint	Ben Chall	Joint	Ben Chall	Joint	Ben Chall	Joint
11–20	$\frac{38}{172}$	$\frac{27}{172}$	$\frac{37}{185}$	$\frac{24}{185}$	$\frac{13}{49}$	$\frac{12}{49}$	$\frac{12}{50}$	$\frac{11}{50}$
%	22.1	15.7	20	13.0	26.5	24.5	24	22.0
21–30	$\frac{18}{179}$	$\frac{13}{179}$	$\frac{31}{189}$	$\frac{23}{185}$	$\frac{10}{44}$	$\frac{8}{44}$	$\frac{8}{53}$	$\frac{5}{53}$
%	10.1	7.3	16.8	12.4	22.7	18.2	15.1	9.4
31–40	$\frac{26}{179}$	$\frac{23}{179}$	$\frac{31}{189}$	$\frac{24}{189}$	$\frac{15}{52}$	$\frac{15}{52}$	$\frac{8}{53}$	$\frac{5}{53}$
%	14.5	12.8	16.4	12.7	28.8	28.8	15.1	9.4
41–50	$\frac{24}{178}$	$\frac{18}{178}$	$\frac{24}{192}$	$\frac{19}{192}$	$\frac{16}{55}$	$\frac{15}{55}$	$\frac{5}{46}$	$\frac{3}{46}$
%	13.5	10.1	12.5	9.9	29.1	27.3	10.9	6.5
21–50	$\frac{68}{536}$	$\frac{54}{536}$	$\frac{86}{566}$	$\frac{66}{566}$	$\frac{41}{151}$	$\frac{38}{151}$	$\frac{21}{152}$	$\frac{13}{152}$
%	12.7	10.1	15.2	11.7	27.2	25.2	13.8	6.5

Table 6.2: Resistance Rates Given Any Signal

couple high rate sessions, and a rank order test on sessions shows the difference to be not significant ($p = 0.2148$, one-sided Mann-Whitney test) .

Result 3. The rate of joint resistance in B are not significantly different from the rates in LC or in SC (when the beneficiary is the free subordinate).

Overall, then, the results are *not* supportive of the expectation-building hypothesis. We introduced the hypothesis that the high rate of beneficiaries challenging — and high rate of coordinated joint resistance — observed in the CR game could be due to strategic decision-making by the subordinates. Thus our results provide support for the hypothesis that beneficiaries who challenge a leader's transgression are exhibiting solidarity with the victim.

6.8 Conclusions

The coordinated resistance game provides a fruitful framework in which to explore a range of motivations for the observed solidarity among subordinates. In our experiment, subjects repeatedly play the CR game under random and anonymous

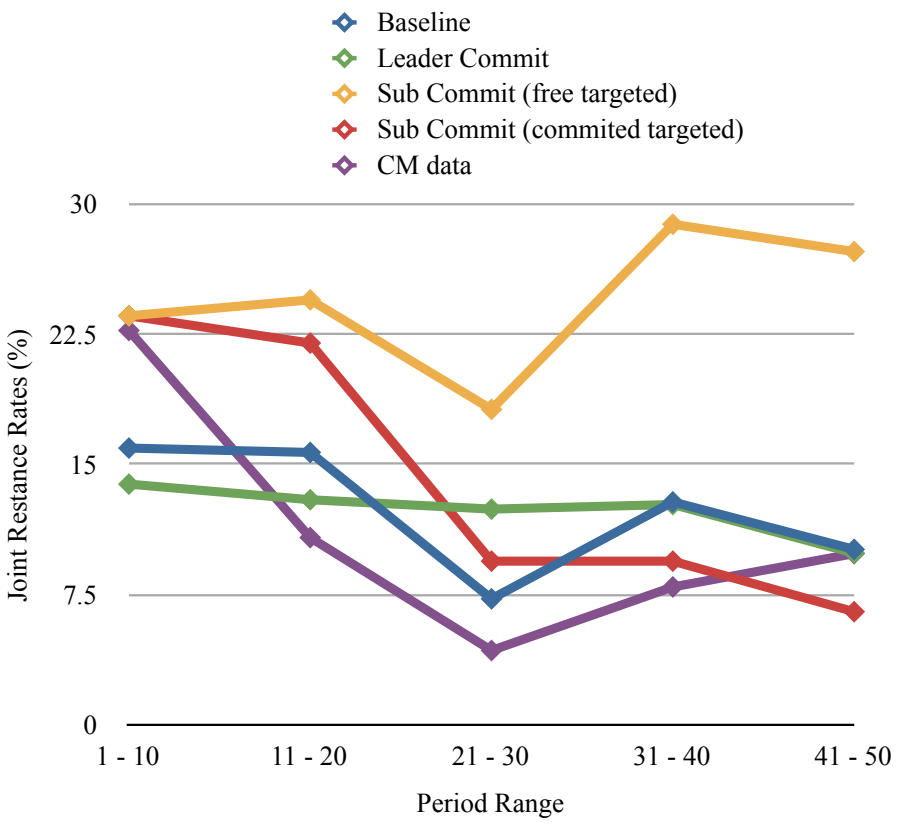


Figure 6.2: Joint Resistance Rates Across Treatments

re-matching and with different types of pre-commitment. We investigate the extent to which the observed coordinated joint resistance — which is not predicted by game theoretic analysis of the game — can be explained by several potential motivations. One potential explanation appeals to norms for fairness: when facing a leader attempting to divide-and-conquer, the subordinate who would benefit by choosing to acquiesce, instead chooses to challenge due to fairness considerations. Another explanation appeals to reputation-building behavior: there is an individual incentive for building a group reputation in an effort to achieve more efficient outcomes in the future. Overall our results provide additional support for the hypothesis that beneficiaries who challenge are exhibiting solidarity with the victim, rather than attempting to strategically expectation build.

References

Boyd, R., H. Gintis, S. Bowles, and P. Richerson (2003), The evolution of altrusitic punishment, *Proceedings of the National Academy of Sciences*, *100*(6), 3531–3535.

Cason, T. N., and V.-L. Mui (2007), Communication and coordination in the laboratory collective resistance game, *Experimental Economics*, *10*(3), 251–267.

Cason, T. N., and V.-L. Mui (2009), Coordinating collective resistance through communication and repeated interaction, working Paper.

Croucher, S. L. (2004), *Globalization and Belonging: The Politics of Identity in a Changing World*, Rowman & Littlefield, Lanham, MD.

DeQuervain, D., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr (2004), The neural basis of altruistic punishment, *Science*, *305*(5688), 1254–1258.

Egas, M., and A. Riedl (2008), The economics of altruistic punishment and the maintenance of cooperation, *Proceedings of the Royal Society B*, *275*(1637), 871–878.

Fehr, E., and U. Fischbacher (2003), The nature of human altruism, *Nature*, (425), 785—791.

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

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

Fehr, E., and S. Gächter (2002), Altruistic punishment in humans, *Nature*, *415*, 137–140.

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

Forsythe, R., J. L. Horowitz, N. E. Savin, and M. Sefton (1994), Fairness in simple bargaining experiments, *Games and Economic Behavior*, *6*(3), 347–369.

Gintis, H., S. Bowles, R. Boyd, and E. Fehr (Eds.) (2005), *Moral Sentiments and Material Interests: Origins, Evidence, and Consequences*, MIT Press, Cambridge, MA.

Gordon, J., and R. A. Lenhardt (2007), Conflict and solidarity between african american and latino immigrant workers, Report filed for The Chief Justice Earl Warren Institute on Race, Ethnicity, and Diversity University of California, Berkeley Law School.

Güth, W., R. Schmittberger, and B. Schwarze (1985), An experimental analysis of ultimatum bargaining, *Journal of Economic Behavior and Organization*, *3*, 367–388.

Horowitz, R. (1997), *'Negro and White, Unite and Fight!' A Social History of Industrial Unionism in Meatpacking, 1930–90*, University of Illinois Press, Champaign–Urbana.

LeDuff, C. (2000), At a slaughterhouse, some things never die: Who kills, who cuts, who bosses can depend on race, New York Times, June 16, 2000.

Ostrom, E. (1990), *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge University Press, New York.

Oyogoa, F. (2009), Conquered then divided on the high seas: International working class solidarity in the vacation cruise ship industry, working Paper.

Rigdon, M. L. (2009), Trust and reciprocity in incentive contracting, *Journal of Economic Behavior and Organization*, *70*, 93–105.

Tasini, J. (2008), The wal-mart divide-and-conquer strategy, Huffington Post, January 25, 2008.

Weingast, B. R. (1997), The political foundations of democracy and the rule of law, *American Political Science Review*, *91*(2), 245–263.

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

Zeitlin, M., and L. F. Weyher (2001), Black and white, unite and fight: Interracial working-class solidarity and racial employment equality, *American Journal of Sociology*, *107*(2), 430–467.

6.9 Appendix 6.A

Instructions for Phase 1

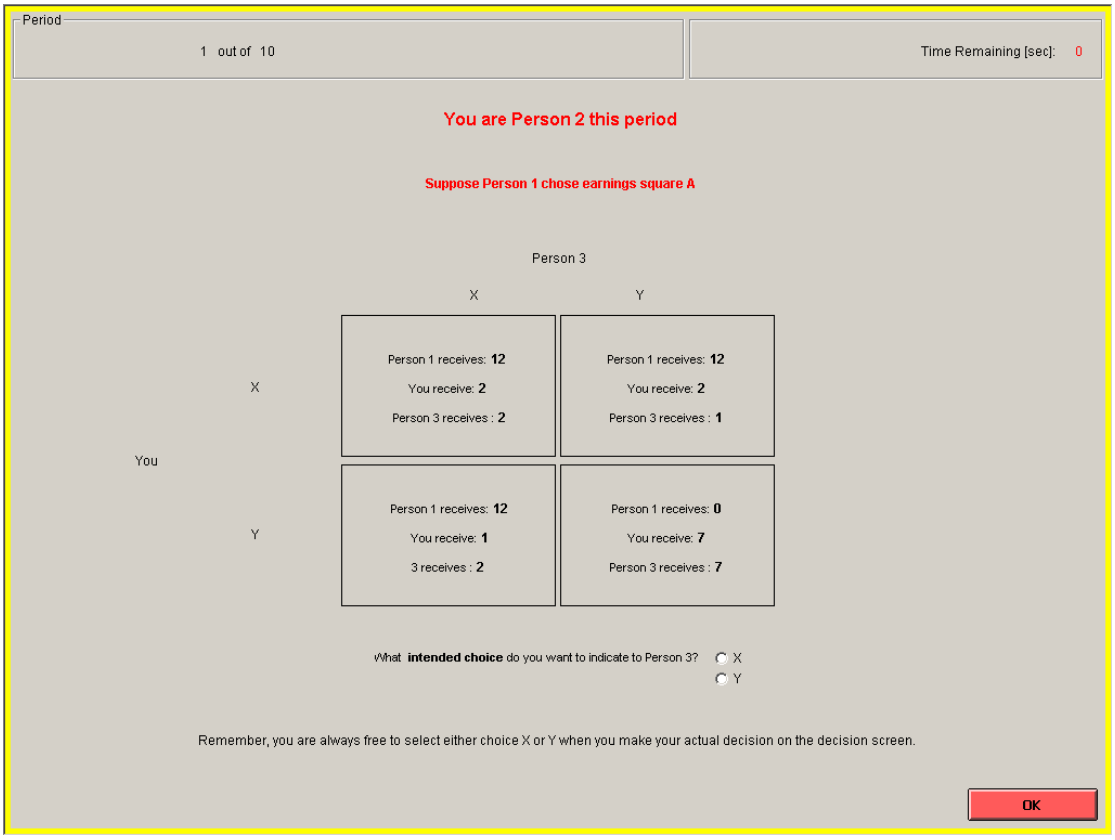
There are two types of decision-making experiments: psychology and economics. In psychology experiments, sometimes the researchers deceive participants involved in the study. When this happens, they are required, before the end of the experiment, to debrief everyone about the nature of the deception. Deception is not permitted in an economics experiment. This is an economics experiment.

Please read the following instructions carefully. If you have a question at any time please raise your hand and an experimenter will come by to answer your question.

Instructions for Phase 1

This is an experiment in the economics of multi-person strategic decision-making. If you follow the instructions and make appropriate decisions, you can earn an appreciable amount of money. The currency used in the experiment is points. Your points will be converted to U.S. Dollars at a rate of $0.09 dollars to one point. At the end of todays session, you will be paid in private and in cash. It is important that you remain silent and do not look at other peoples work. If you have any questions, or need assistance of any kind, please raise your hand and an experimenter will come to you. If you talk, laugh, exclaim out loud, etc., you will be asked to leave and you will not be paid. We expect and appreciate your cooperation. There will be **two phases** to the experiment: Phase 1 and Phase 2. Phase 1 will consist of 10 periods. The 18 participants in todays experiment will be randomly split each period between three equal-sized groups, designated as **Person 1**, **Person 2** and **Person 3** groups. If you are designated as a Person 1, then you remain in this same role throughout both phases of the experiment. Participants who are not designated as a Person 1 switch randomly between the Person 2 and Person 3 roles in different

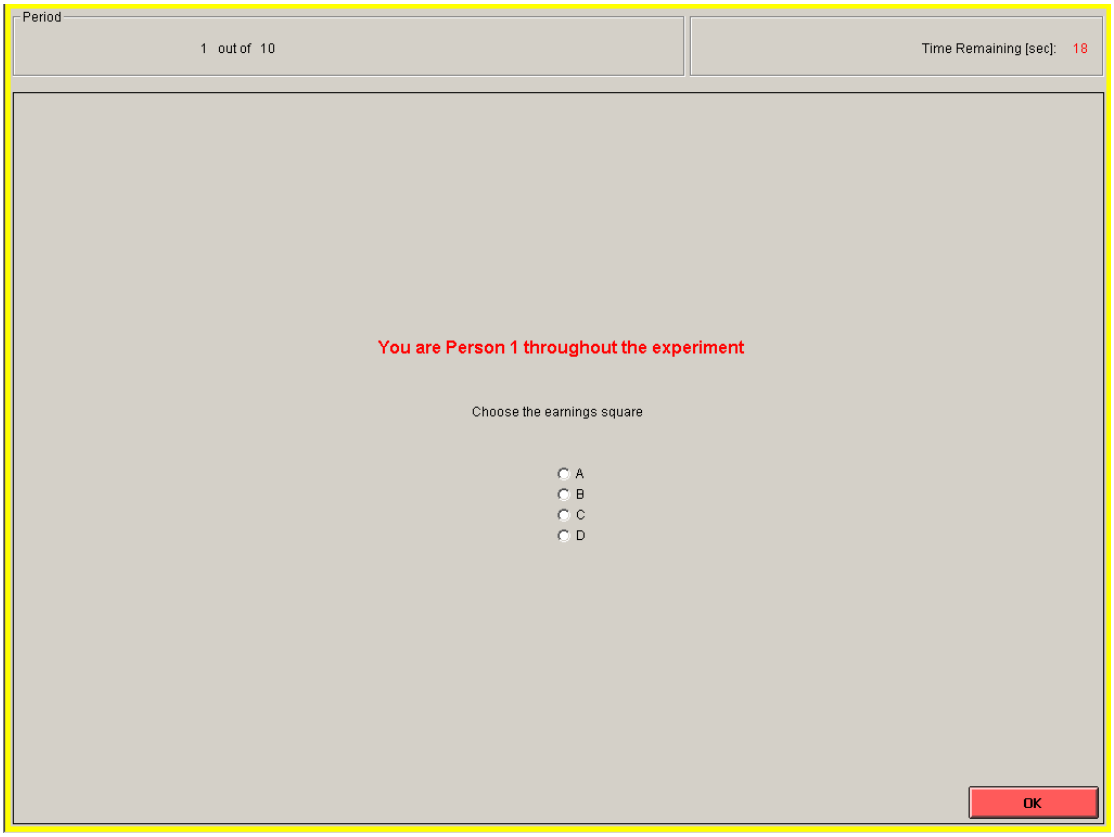
decision-making periods throughout both phases of the experiment. Phase 2 will be a similar decision-making task, and you will have the same role in it that you have in Phase 1. Further instructions will be provided before Phase 2 begins. At the beginning of each decision-making period you will be randomly re-grouped with two other participants to form a three-person group, with one person of each type in each group. The groupings change every period, since you will be randomly re-grouped in each and every period.



Your Choice During each period, you and all other participants will make one choice. Earnings tables are provided on separate papers, which tell you the earnings you receive given the choices that you and others in your group make. If you are **Person 1** then you choose the earnings square, either **A**, **B**, **C** or **D**. You make this choice before the other two people in your group make their choices, on a decision screen as shown on page 4. While **Person 1** chooses the earnings square, however,

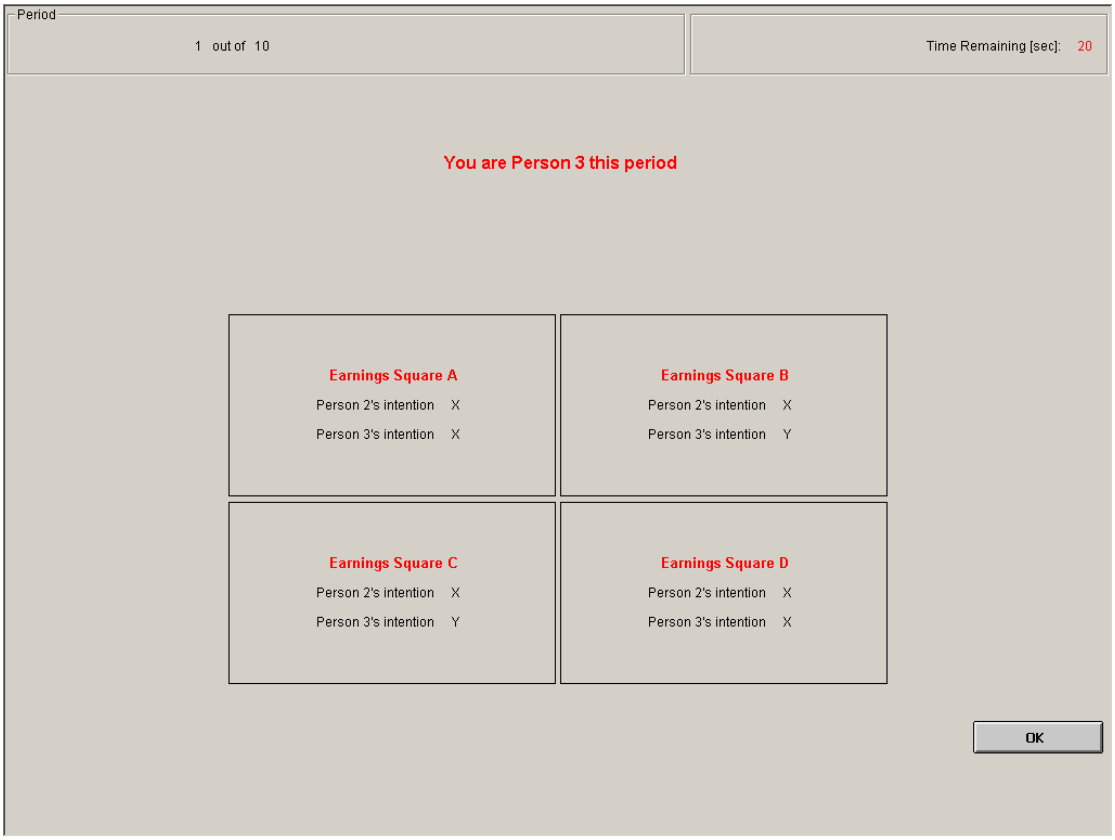
Persons 2 and 3 have an opportunity to communicate to each other an intended choice for every one of the four possible earnings squares. Persons 2 and 3 indicate their intended choices simultaneously; for example, if you are Person 3 then you do not learn the intended choices of Person 2 until after you indicate all your intended choices. As noted on the example Intention Screen for Person 2 below, Persons 2 and 3 are not required to make the same actual choice as corresponding to their intended choice, and they are always free to select either choice X or Y when they make their actual decision.

While Persons 2 and 3 are indicating their intended choices, Person 1 chooses the earnings square using the Person 1 Decision Screen as shown below.

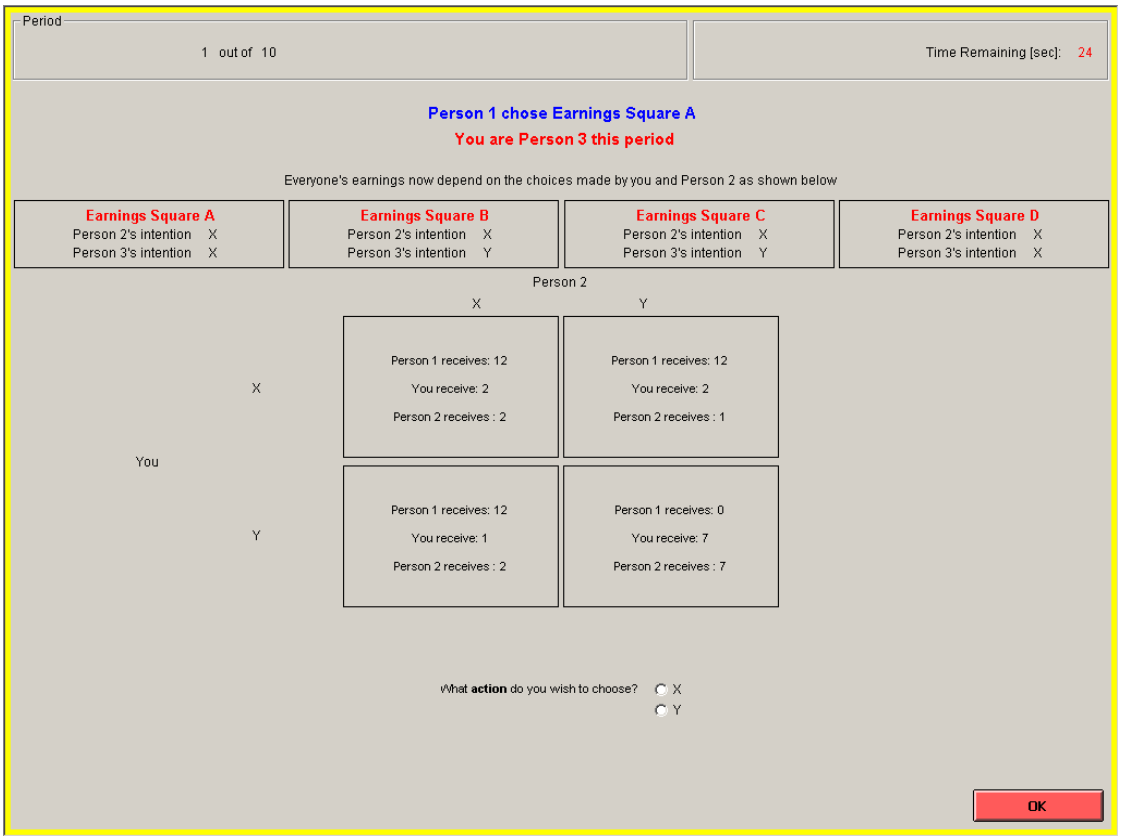


After both Persons 2 and 3 have finished indicating their intended choices, the computer program displays all the intended choices to both Person 2 and Person 3 as shown on the next page (page 5). The intentions and Person 1s earning square

choice are also shown on the Decision Screen for Persons 2 and 3, as shown on page 6. Persons 2 and 3 then make their actual choice simultaneously; for example, if you are Person 2 then you do not learn the actual choice of Person 3 until after you make your choice. Both Persons 2 and 3 may choose either **X** or **Y**.



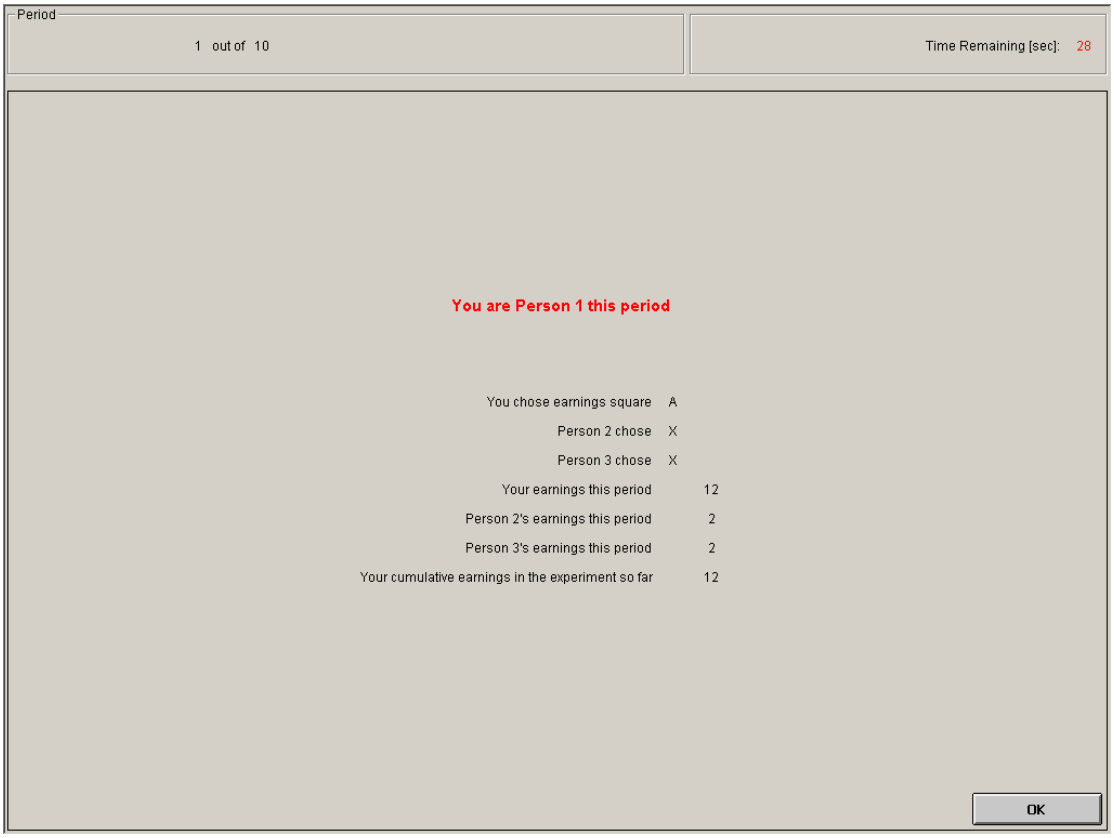
Your earnings from the choices each period are found in the box determined by you and the other two people that you are grouped with for the current decision making period. If both Persons 2 and 3 choose **X**, then earnings are paid as shown in the box in the upper left on the screen. If both Persons 2 and 3 choose **Y**, then earnings are paid as shown in the box in the lower right on the screen. The other two boxes indicate earnings when one chooses **X** and the other chooses **Y**. To illustrate with a random example: if Person 1 chooses earnings square **A**, Person 2 chooses **X** and Person 3 chooses **Y**, then Person 1 earns 12, Person 2 earns 2, and Person 3 earns 1. You can find these amounts by looking at the appropriate square and box in



your page of earnings tables. In summary, Persons 2 and 3 indicate simultaneously their intended choice for each of the four earning squares that Person 1 can choose, while Person 1 chooses the earnings square. When both Persons 2 and 3 have finished indicating their intentions, the computer program displays all the intended choices to both Person 2 and Person 3. The computer program also displays the earnings square chosen by Person 1. Persons 2 and 3 then simultaneously make their choices of X and Y. Remember, Persons 2 and 3 are not required to make the same actual choice corresponding to their intended choice, and they are always free to select either choice X or Y when they make their actual decision.

The End of the Period After everyone has made choices for the current period you will be automatically switched to the outcome screen, as shown below. This screen displays your choice as well as the choices of the people you are grouped with for the current decision making period. It also shows your earnings for this period and your

earnings for the experiment so far.



Once the outcome screen is displayed you should record your choice and the choice of the others in your group on your Personal Record Sheet. Also record your current and cumulative earnings. Then click on the *OK* button on the lower right of your screen. Remember, at the start of the next period all participants are randomly re-grouped, and you are randomly re-grouped each and every period of the experiment. We will now pass out a questionnaire to make sure that all participants understand how to read the earnings tables and understand other important features of these instructions. Please fill it out now. Raise your hand when you are finished and we will collect it. If there are any mistakes on any questionnaire, I will summarize the relevant part of the instructions again. Do not put your name on the questionnaire.

6.10 Appendix 6.B

Instructions for Phase 2: Baseline

We are entering the second and final phase of the experiment.

This phase will last for 40 periods.

Each of these 40 periods is identical to those in Phase 1.

Please continue to record all decisions on the Personal Record Sheet for your role; they are attached to Phase 2 instructions.

6.11 Appendix 6.C

Instructions for Phase 2: Leader Commitment

We are entering the second and final phase of the experiment.

This phase will last for 40 periods.

The only difference between periods in this phase and periods in the previous phase is that each Person 1 will choose **all of their actions** for the next 40 periods now, before any more periods are played.

That means in the next 40 periods, each Person 1 will be pre-committed to actions that they will specify for each period after they finish reading this screen.

Those in the roles of Person 2 and Person 3 will make choices identically to how they did in Phase 1.

Persons 2 and 3 will now have a few minute wait while those in the role of Person 1 make their next 40 choices.

Please continue to record all decisions on the Personal Record Sheet for your role; they are attached to Phase 2 instructions.

6.12 Appendix 6.D

Instructions for Phase 2: Both Commitment

We are entering the second and final phase of the experiment.

This phase will last for 40 periods.

There are **two important differences** between periods in this phase and periods in the previous phase:

1. All Persons 1 will choose **all of their actions** for the next 40 periods now, before any more periods are played.

   That means in the next 40 periods, each Person 1 will be pre-committed to actions that they will specify for each period after they finish reading this.

2. Half of those in the roles of Person 2 and Person 3 will indicate their **intended choices and choose all of their actions for all periods without observing any choices made by either Person 1 or Person 2 and 3.**

   That means in the next 40 periods, half of the Persons 2 and 3 will be pre-committed to intended choices and actions that they will choose before any more periods occur.

The other half of the Persons 2 and Person 3 will make their choices in each period as in Phase 1.

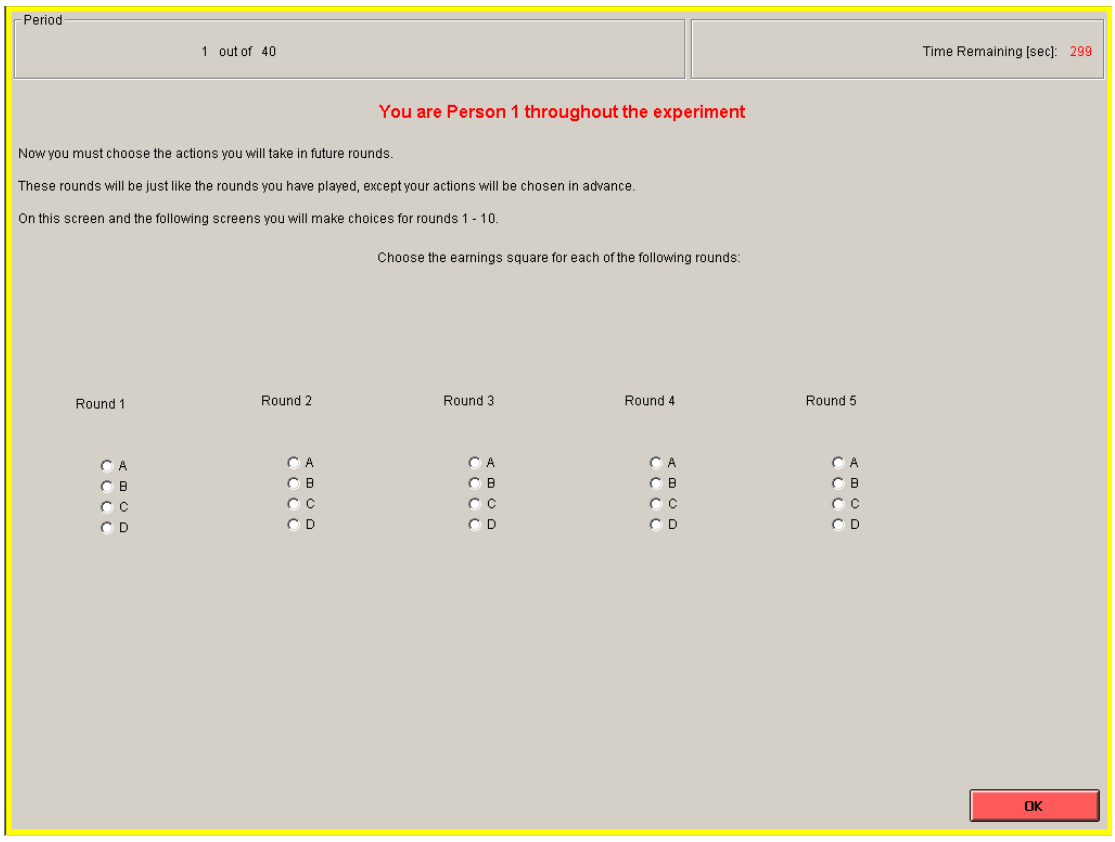
Note: for earnings square D, all Persons 2 and 3 will be restricted to intended choice "X" and action "X".

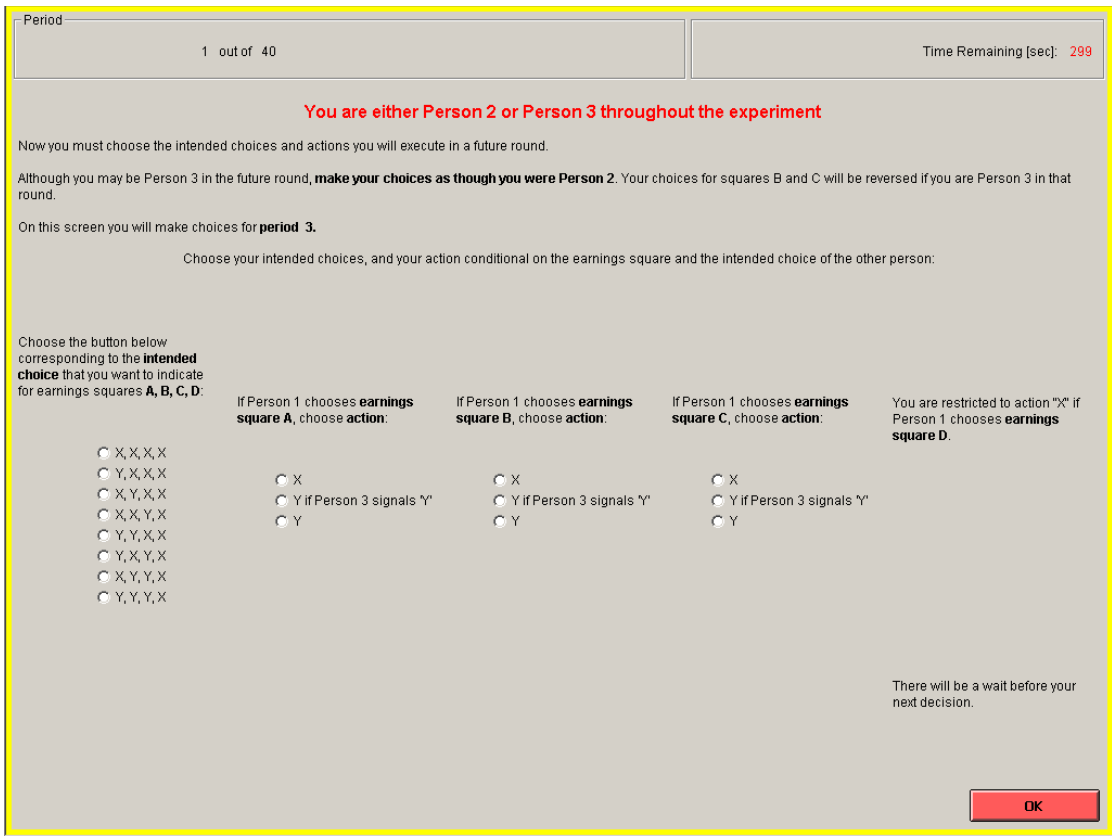
The division of the Persons 2 and 3 into these two categories will remain constant across all 40 periods of this phase.

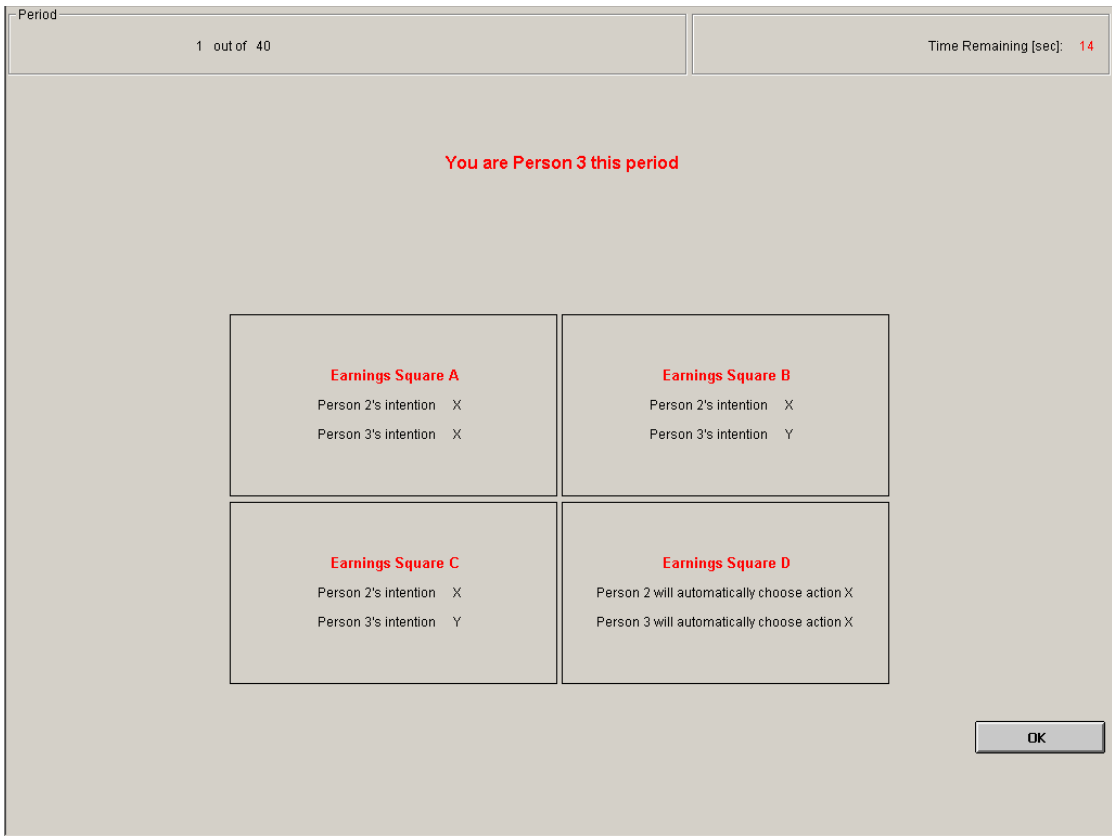
All participants will continue to be randomly re-grouped in each and every period.

Additionally, in each and every period, each Person 1 will be grouped with **one** Person 2 or 3 who is **pre-committed**, i.e. has already indicated an intended choice

and chosen an action for the period without observing any choices by others, and **one** Person 2 or 3 who is **not pre-committed**, i.e. indicates his/her intended choice and action for each period as in Phase 1.


Period
1 out of 40
Time Remaining [sec]: 299
You are Person 1 throughout the experiment
Now you must choose the actions you will take in future rounds.
These rounds will be just like the rounds you have played, except your actions will be chosen in advance.
On this screen and the following screens you will make choices for rounds 1 - 10.
Choose the earnings square for each of the following rounds:
Round 1
Round 2
Round 3
Round 4
Round 5
A
B
C
D
OK


Period
1 out of 40
Time Remaining [sec]: 299
You are either Person 2 or Person 3 throughout the experiment
Now you must choose the intended choices and actions you will execute in a future round.
Although you may be Person 3 in the future round, make your choices as though you were Person 2. Your choices for squares B and C will be reversed if you are Person 3 in that round.
On this screen you will make choices for period 3.
Choose your intended choices, and your action conditional on the earnings square and the intended choice of the other person:
Choose the button below corresponding to the intended choice that you want to indicate for earnings squares A, B, C, D:
X, X, X, X
Y, X, X, X
X, Y, X, X
X, X, Y, X
Y, Y, X, X
Y, X, Y, X
X, Y, Y, X
Y, Y, Y, X
If Person 1 chooses earnings square A, choose action:
X
Y if Person 3 signals 'Y'
Y
If Person 1 chooses earnings square B, choose action:
X
Y if Person 3 signals 'Y'
Y
If Person 1 chooses earnings square C, choose action:
X
Y if Person 3 signals 'Y'
Y
You are restricted to action "X" if Person 1 chooses earnings square D.
There will be a wait before your next decision.
OK


Period
1 out of 40
Time Remaining [sec]: 14
You are Person 3 this period
Earnings Square A
Person 2's intention X
Person 3's intention X
Earnings Square B
Person 2's intention X
Person 3's intention Y
Earnings Square C
Person 2's intention X
Person 3's intention Y
Earnings Square D
Person 2 will automatically choose action X
Person 3 will automatically choose action X
OK