ROBUST IMPLEMENTATION IN GENERAL MECHANISMS

By

Dirk Bergemann and Stephen Morris

June 2008

COWLES FOUNDATION DISCUSSION PAPER NO. 1666

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Robust Implementation in General Mechanisms*

Dirk Bergemann† Stephen Morris‡

June 2008

Abstract

A social choice function is *robustly implemented* if every equilibrium on every type space achieves outcomes consistent with it. We identify a *robust monotonicity* condition that is necessary and (with mild extra assumptions) sufficient for robust implementation.

Robust monotonicity is strictly stronger than both Maskin monotonicity (necessary and almost sufficient for complete information implementation) and ex post monotonicity (necessary and almost sufficient for ex post implementation). It is equivalent to Bayesian monotonicity on all type spaces.

KEYWORDS: Mechanism Design, Implementation, Robustness, Common Knowledge, Interim Equilibrium, Dominant Strategies.

JEL CLASSIFICATION: C79, D82

*This research is supported by NSF Grants #CNS-0428422 and #SES-0518929. The first author gratefully acknowledges support through a DFG Mercator Research Professorship at the Center of Economic Studies at the University of Munich. We thank Matt Jackson and Andy Postlewaite for helpful discussions and Richard van Weelden for research assistance. This paper supersedes and incorporates results reported earlier in Bergemann and Morris (2005a).

†Department of Economics, Yale University, New Haven, CT 06511, dirk.bergemann@yale.edu

‡Department of Economics, Princeton University, Princeton, NJ 08544, smorris@princeton.edu

1 Introduction

The objective of mechanism design is to construct mechanisms (or game forms) such that privately informed agents have an incentive to reveal their information to a principal who seeks to realize a social choice function. The revelation principle establishes that if any mechanism can induce the agents to report their information, then the agents will also have an incentive to report truthfully in the direct mechanism. Given the beliefs of the agents, the truthtelling constraints then reduce, in the direct mechanism,- to the Bayesian incentive compatibility conditions.

There are two important limitations of Bayesian incentive compatibility analysis. First, the analysis typically assumes a commonly known common prior over the agents' types. This assumption may be too stringent in practice. In the spirit of the "Wilson doctrine" (Wilson (1987)), we would like implementation results that are *robust* to different assumptions about what agents do or do not know about other agents' types. Second, the revelation principle only establishes that the direct mechanism has *an* equilibrium that achieves the social choice function. In general, there may be other equilibria that deliver undesirable outcomes. We would like to achieve *full* implementation, i.e., show the existence of a mechanism all of whose equilibria deliver the social choice function. We studied the first "robustness" problem in an earlier work, Bergemann and Morris (2005b). The second "full implementation" problem has been the subject of a large literature. In the incomplete information context, key full implementation references are Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989) and Jackson (1991). In this paper, we study "robust implementation" where we require robustness and full implementation simultaneously.

Interim implementation on all type spaces is possible if and only if it is possible to implement the social choice function using an iterative deletion procedure. We refer to the resulting notion as *rationalizable implementation*. We fix a mechanism and iteratively delete messages for each payoff type that are strictly dominated by another message for each payoff type profile and message profile that has survived the procedure. This observation about iterative deletion illustrates a general point well-known from the literature on epistemic foundations of game theory (e.g., Brandenburger and Dekel (1987), Battigalli and Siniscalchi (2003)): equilibrium solution concepts only have bite if we make strong assumptions about type spaces, i.e., we assume small type spaces where the common prior assumption holds.

We exploit this equivalence between robust and rationalizable implementation to obtain necessary and sufficient conditions for robust implementation in general environments. Our necessity argument is conceptually novel, exploiting the iterative characterization. The necessary conditions for robust implementation are ex post incentive compatibility of the social choice function and a condition - *robust monotonicity* - that is equivalent to requiring Bayesian monotonicity on every

type space. Suppose that we fix a "deception" specifying, for each payoff type θ_i of each agent, a set of types that he might misreport himself to be. We require that for some agent i and a type misreport of agent i under the deception, for every misreport θ'_{-i} that the other agents might make under the deception, there exists an outcome y which is strictly preferred by agent i to the outcome he would receive under the social choice function for *every* possible payoff type profile that might misreport θ'_{-i}; where this outcome y satisfies the extra restriction that no payoff type of agent i prefers outcome y to the social choice function if the other agents were really types θ'_{-i}. This condition - while a little convoluted - is easier to interpret than the interim (Bayesian) monotonicity conditions.

The sufficiency argument requires only a modest strengthening of the necessary condition by guaranteeing that the preference profile of each agent satisfies a (conditional) no total indifference property. Under this no total indifference property, we show that the necessary conditions are also sufficient for robust implementation. The sufficient conditions guarantee robust implementation in pure, but more generally also in mixed strategies. Our robust analysis thus removes the frequent gap between pure and mixed strategy implementation in the literature.

In this paper, we follow the classic implementation literature in allowing for arbitrary mechanisms, including modulo and integer games. By allowing for these mechanisms, we are able to make tight connections with the existing implementation literature. Allowing for these badly behaved mechanisms does complicate our analysis: for example, we must allow for transfinite iterated deletion of best responses in our definition of rationalizable implementation. Given the complications arising from infinite mechanisms, we report new necessary conditions for robust implementation in the context of finite mechanisms. We also report how our earlier research can be used to show that these necessary conditions are sufficient conditions for finite mechanisms either in well-behaved, but restricted, environments (Bergemann and Morris (2007)) or under a virtual rather than exact implementation requirement (Bergemann and Morris (2008c)).

Our results extend the classic literature on Bayesian implementation due to Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989) and Jackson (1991). We focus in this paper on an indirect approach to extending these results. We first note the equivalence between robust implementation and rationalizable implementation. We then exploit the equivalence to report a direct argument showing that robust monotonicity is a necessary and almost sufficient condition for rationalizable implementation. But in the light of the classic literature, we know that a necessary and almost sufficient condition for robust implementation must be Bayesian monotonicity on all type spaces. We confirm and clarify our results by directly checking that robust monotonicity is equivalent to Bayesian (or interim) monotonicity on all type spaces. Figure 1 gives a stylized account of the connection between these alternative approaches.

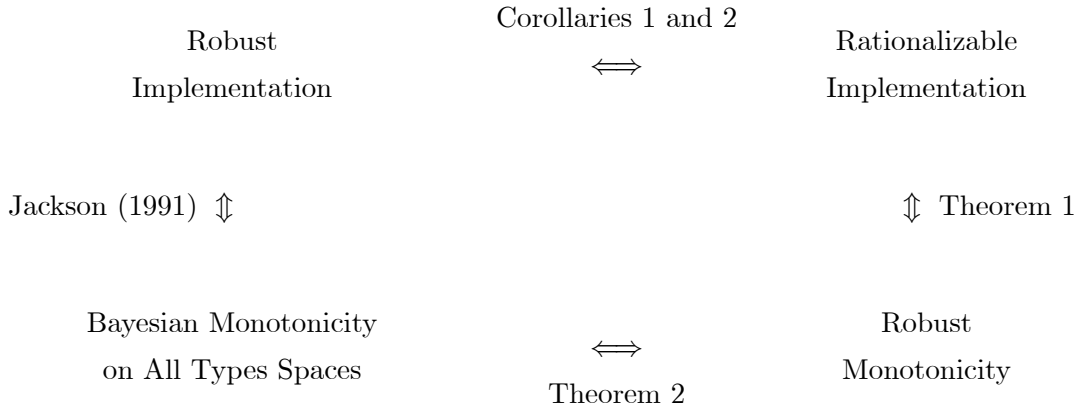

Figure 1: Relationship between Bayesian and Robust Implementation / Monotonicity

In the implementation literature, it is a standard practice to obtain the sufficiency results with augmented mechanisms. By augmenting the direct mechanism with additional messages, the designer may elicit additional information about undesirable equilibrium play by the agents. Yet, in many applied economic settings, single crossing or supermodular preference assumptions allow direct implementation. In a companion paper, Bergemann and Morris (2007), we provide necessary and sufficient conditions for robust implementation in the direct mechanism. The main results of this paper apply to environments where each agent's type profile can be aggregated into a one dimensional sufficient statistic for each player and where the preferences are single crossing with respect to this statistic. These restrictions incorporate many economic models with interdependence in the literature. We show that besides an incentive compatibility condition, in this case the strict ex post incentive compatibility condition, a *contraction* property which requires that there is *not too much* interdependence in agents' types, together present necessary and sufficient conditions for robust implementation in direct mechanisms.

The robust monotonicity condition is stronger than both the Maskin and the Bayesian monotonicity conditions. In the context of robust implementation, it is then natural to ask whether a relaxation from the exact to the virtual implementation condition may lead to more permissive results. In Bergemann and Morris (2008c) we characterize the necessary and sufficient conditions for *robust virtual* implementation. There we show that a social choice function can be *robustly virtually implemented* if and only if the social choice function is ex post incentive compatible and robust measurable. In this contribution, we note that robust measurability remains a necessary condition for robust (exact) implementation, but it is not sufficient anymore.

The results in this paper concern full implementation. An earlier paper of ours, Bergemann and

Morris (2005b), addresses the analogous questions of robustness to rich type spaces, but looking at the question of partial implementation, i.e., does there exist a mechanism such that *some* equilibrium implements the social choice function. We showed that ex post (partial) implementation of the social choice function is a necessary and sufficient condition for partial implementation on all type spaces.[1] This paper establishes that an analogous result does *not* hold for full implementation.

In a related paper, Bergemann and Morris (2008a), we therefore investigate the notion of ex post implementation. The necessary and sufficient conditions there straddle the implementation conditions for Nash and Bayesian-Nash respectively, as an ex post equilibrium is a Nash equilibrium at every incomplete information (Bayesian) type profile. However in contrast to the iterative argument pursued here, the basic reasoning in Bergemann and Morris (2008a) invokes more traditional equilibrium arguments. By comparing the conditions for ex post and robust implementation, it becomes apparent that robust implementation typically imposes additional constraints on the allocation problem. In Bergemann and Morris (2008a), we showed that in single crossing environments, the same single crossing conditions which guarantee incentive compatibility also guarantee full implementation. In contrast, in the aggregation environment discussed above, we show that robust implementation imposes a strict bound on the interdependence of the preferences, which is not required by the truthtelling conditions. A contraction mapping behind the iterative argument directly points to the source of the restriction of the interaction term.

The remainder of the paper is organized as follows. Section 2 describes the formal environment and solution concepts. Section 3 establishes necessary conditions for robust implementation in finite mechanisms. In addition, we present restrictions on the environment and weaker implementation notions under which the necessary conditions are also sufficient conditions. Section 4 establishes the relation between rationalizable and robust implementation in infinite mechanisms. Section 5 reports our main result on the necessary and sufficient conditions for robust implementation. Section 6 discusses extensions and variations of our implementation results, examining the role of lotteries and pure strategies and the relationship with Nash equilibrium and ex post equilibrium implementation. The appendix contains some additional examples.

[1] This result does not extend to social choice correspondences.

2 Setup

2.1 The Payoff Environment

We consider a finite set of agents, $1, 2, ..., I$. Agent i's *payoff type* is $\theta_i \in \Theta_i$. We write $\theta \in \Theta = \Theta_1 \times \cdots \times \Theta_I$. There is a set of outcomes Z. We assume that each Θ_i and Z are countable.[2] Each individual has a von Neumann Morgenstern utility function $u_i : Z \times \Theta \to \mathbb{R}$. Thus we are in the world of interdependent types, where an agent's utility depends on other agents' payoff types. We allow for lotteries over deterministic outcomes.[3] Let $Y \triangleq \Delta(Z)$ and extend u_i to the domain $Y \times \Theta$ in the usual way:

$$u_i(y, \theta) \triangleq \sum_{z \in Z} y(z) u_i(z, \theta).$$

A social choice function is a mapping $f : \Theta \to Y$. If the true payoff type profile is θ, the planner would like the outcome to be $f(\theta)$. In this paper, we restrict our analysis to the implementation of a social choice function rather than a social choice correspondence or set.[4]

2.2 Type Spaces

We are interested in analyzing behavior in a variety of type spaces, many of them with a richer set of types than payoff types. For this purpose, we shall refer to agent i's *type* as $t_i \in T_i$, where T_i is a countable set. A type of agent i must include a description of his payoff type. Thus there is a function $\widehat{\theta}_i : T_i \to \Theta_i$ with $\widehat{\theta}_i(t_i)$ being agent i's payoff type when his type is t_i. A type of agent i must also include a description of his beliefs about the types of the other agents; thus there is a function $\widehat{\pi}_i : T_i \to \Delta(T_{-i})$ with $\widehat{\pi}_i(t_i)$ being agent i's *belief type* when his type is t_i. Thus $\widehat{\pi}_i(t_{-i})[t_i]$ is the probability that type t_i of agent i assigns to other agents having types t_{-i}. A *type space* is a collection:

$$\mathcal{T} = \left(T_i, \widehat{\theta}_i, \widehat{\pi}_i\right)_{i=1}^{I}.$$

2.3 Mechanisms

A planner must choose a *game form* or *mechanism* for the agents to play in order to determine the social outcome. Let M_i be the countably infinite set of messages available to agent i. We denote

[2] The countable types restriction clarifies the relation to the existing literature. We postpone until Section 6.3 a discussion of what happens if we allow for uncountable payoff types, types and pure outcomes.

[3] The role of the lottery assumption and what happens when we drop it are discussed in Section 6.1.

[4] One reason why the extension to social choice correspondences is not straightforward is that, with social choice correspondences, the incentive compatibility conditions that arise from requiring partial implementation are typically weaker than ex post incentive compatibility, as shown by examples in Bergemann and Morris (2005b).

the generic message by $m_i \in M_i$ and let $m \in M = M_1 \times \cdots \times M_I$. Let $g(m)$ be the distribution over outcomes if action profile m is chosen. Thus a mechanism is a collection

$$\mathcal{M} = (M_1, ..., M_I, g(\cdot)),$$

where $g : M \to Y$.

2.4 Solution Concepts

Now holding fixed the payoff environment, we can combine a type space $\mathcal{T}$ with a mechanism $\mathcal{M}$ to get an incomplete information game $(\mathcal{T}, \mathcal{M})$. The payoff of agent i if message profile m is chosen and type profile t is realized is then given by

$$u_i\left(g(m), \widehat{\theta}(t)\right).$$

A pure strategy for agent i in the incomplete information game $(\mathcal{T}, \mathcal{M})$ is given by

$$s_i : T_i \to M_i.$$

A (behavioral) strategy is given by

$$\sigma_i : T_i \to \Delta(M_i).$$

The objective of this paper is to obtain implementation results for interim, or Bayesian Nash, equilibria on all possible types spaces.[5] The notion of interim equilibrium for a given type space $\mathcal{T}$ is defined in the usual way.

Definition 1 (Interim equilibrium)
A strategy profile $\sigma = (\sigma_1, ..., \sigma_I)$ is an interim equilibrium of the game $(\mathcal{T}, \mathcal{M})$ if, for all i, t_i and m_i with $\sigma_i(m_i|t_i) > 0$,

$$\sum_{t_{-i} \in T_{-i}} \sum_{m_{-i} \in M_{-i}} \left(\prod_{j \neq i} \sigma_j(m_j|t_j) \right) u_i\left(g(m_i, m_{-i}), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_{-i})[t_i]$$

$$\geq \sum_{t_{-i} \in T_{-i}} \sum_{m_{-i} \in M_{-i}} \left(\prod_{j \neq i} \sigma_j(m_j|t_j) \right) u_i\left(g(m_i', m_{-i}), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_{-i})[t_i]$$

for all m_i'.

[5] We label these "interim" equilibria rather than "Bayesian" equilibria in light of the fact that our type space does not necessarily have a common prior.

Requiring "robust" implementation, i.e., for "all type spaces", will push the solution concept in the direction of rationalizability. Consequently we define a message correspondence profile $S = (S_1, ..., S_I)$, where each

$$S_i : \Theta_i \to 2^{M_i} \tag{1}$$

and we write $\mathcal{S}$ for the collection of message correspondence profiles. The collection $\mathcal{S}$ is a lattice with the natural ordering of set inclusion: $S \leq S'$ if $S_i(\theta_i) \subseteq S'_i(\theta_i)$ for all i and θ_i. The largest element is $\overline{S} = \left(\overline{S}_1, ..., \overline{S}_I\right)$, where $\overline{S}_i(\theta_i) = M_i$ for each i and θ_i. The smallest element is $\underline{S} = (\underline{S}_1, ..., \underline{S}_I)$, where $\underline{S}_i(\theta_i) = \varnothing$ for each i and θ_i.

We define an operator b to iteratively eliminate never best responses. To this end, we denote the belief of agent i over message and payoff type profiles of the remaining agents by

$$\lambda_i \in \Delta(M_{-i} \times \Theta_{-i}).$$

The operator $b : \mathcal{S} \to \mathcal{S}$ is now defined as:

$$b_i(S)[\theta_i] = \left\{ m_i \in M_i \;\middle|\; \exists \lambda_i \text{ s.th.:} \begin{array}{ll} (1) & \lambda_i(m_{-i}, \theta_{-i}) > 0 \Rightarrow m_j \in S_j(\theta_j),\ \forall\, j \neq i; \\ (2) & \sum\limits_{m_{-i}, \theta_{-i}} \lambda_i(m_{-i}, \theta_{-i})\, u_i(g(m_i, m_{-i}), (\theta_i, \theta_{-i})) \\ & \geq \\ & \sum\limits_{m_{-i}, \theta_{-i}} \lambda_i(m_{-i}, \theta_{-i})\, u_i(g(m'_i, m_{-i}), (\theta_i, \theta_{-i})),\ \forall m'_i \in M_i; \end{array} \right\}. \tag{2}$$

We observe that b is increasing by definition: i.e., $S \leq S' \Rightarrow b(S) \leq b(S')$. By Tarski's fixed point theorem, there is a largest fixed point of b, which we label $S^{\mathcal{M}}$. Thus (i) $b\left(S^{\mathcal{M}}\right) = S^{\mathcal{M}}$ and (ii) $b(S) = S \Rightarrow S \leq S^{\mathcal{M}}$. We can also construct the fixed point $S^{\mathcal{M}}$ by starting with $\overline{S}$ - the largest element of the lattice - and iteratively applying the operator b. If the message sets and types are finite, we have

$$S_i^{\mathcal{M}}(\theta_i) \triangleq \bigcap_{n \geq 1} b_i\left(b^n\left(\overline{S}\right)\right)[\theta_i].$$

But because the mechanism $\mathcal{M}$ may be infinite, transfinite induction may be necessary to reach the fixed point.[6] It is useful to define

$$S_i^{\mathcal{M},k}(\theta_i) \triangleq b_i\left(b^{k-1}\left(\overline{S}\right)\right)[\theta_i],$$

[6] Lipman (1994) contains a formal description of the transfinite induction required (for the case of complete information, but nothing important changes with incomplete information). As he notes "we remove strategies which are never a best reply, taking limits where needed".

again using transfinite induction if necessary. Thus $S_i^{\mathcal{M}}(\theta_i)$ are the set of messages surviving (transfinite) iterated deletion of never best responses. $S_i^{\mathcal{M}}(\theta_i)$ is the set of messages that type θ_i might send consistent with knowing that his payoff type is θ_i, common knowledge of rationality and the set of possible payoff types of the other players, but no restrictions on his beliefs and higher order beliefs about other types.

If message sets are finite (or compact), a well known duality argument implies that never best responses are equivalent to strictly dominated actions. However, the equivalence does not hold with infinite (non-compact) message sets.[7] In a compact message analysis, Chung and Ely (2001) consider a version of this solution concept in an incomplete information mechanism design context with dominated (not strictly dominated) messages deleted at each round. We observe that the solution concept defined through the iterative application of the operator b is weaker than the notion of interim rationalizability for a given type space $\mathcal{T}$.[8] Under b, every agent i is allowed to hold arbitrary beliefs about Θ_{-i} and is not restricted to a particular posterior distribution over Θ_{-i}. On the other hand, if the type space $\mathcal{T}$ were the universal type space, then $S_i^{\mathcal{M}}(\theta_i)$ would be equal to the union of all interim rationalizable actions of agent i over all types $t_i \in T_i$ whose payoff type profile coincides with θ_i, or $\widehat{\theta}_i(t_i) = \theta_i$. We refer to $S_i^{\mathcal{M}}(\theta_i)$ as the *rationalizable* messages of type θ_i of agent i in mechanism $\mathcal{M}$.

2.5 Implementation

We now define the notions of interim, robust and rationalizable implementation.

Definition 2 (Interim Implementation)
Social choice function f is interim implemented on type space $\mathcal{T}$ by mechanism $\mathcal{M}$ if the game $(\mathcal{T}, \mathcal{M})$ has an equilibrium and every equilibrium σ of the game $(\mathcal{T}, \mathcal{M})$ satisfies

$$\sigma(m|t) > 0 \Rightarrow g(m) = f\left(\widehat{\theta}(t)\right).$$

We note that a tradition in the implementation literature commonly restricts attention to pure strategy equilibria, but we allow mixed strategy equilibria.

[7] The following simple example (suggested to us by Andrew Postlewaite) illustrates the non-equivalence. Players 1 and 2 each choose a non-negative integer, k_1 and k_2 respectively. The payoff to player 1 from $k_1 = 0$ is 1. The payoff to player 1 from action $k_1 \geq 1$ is 2 if $k_1 > k_2$, 0 otherwise. For any belief that player 1 has about 2's actions, there is a (sufficiently high) action from player 1 that gives him a payoff greater than 1. Thus action 0 is never a best response for player 1. However, for any mixed strategy of player 1, there is a (sufficiently high) action of player 2 such that action 0 is a better response for player 1 than the mixed strategy. Thus action 0 is not strictly dominated.

[8] For the notion of interim rationalizability, see Battigalli and Siniscalchi (2003) and Dekel, Fudenberg, and Morris (2007).

Definition 3 (Robust Implementation)
Social choice function f is robustly implemented by mechanism $\mathcal{M}$ if, for every $\mathcal{T}$, f is interim implemented on type space $\mathcal{T}$ by mechanism $\mathcal{M}$. Social choice function f is robustly implementable if there exists a mechanism $\mathcal{M}$ such that f is robustly implemented by mechanism $\mathcal{M}$.

We observe that the notion of robust implementation requires that we can find a mechanism $\mathcal{M}$ which implements f for every type space $\mathcal{T}$. A weaker requirement would be to ask that for every type space $\mathcal{T}$ there exists a, possibly different, mechanism $\mathcal{M}$ such that f is implemented. This weaker notion would still lead to the same necessary condition as the stronger implementation version we pursue here, and we believe that it would not lead to a substantial change in the sufficiency conditions either.

The notion of robust implementation requires that a social choice function f can be interim implemented for all type spaces $\mathcal{T}$. As we look for necessary and sufficient conditions for robust implementation, conceptually there are (at least) two approaches to obtain the conditions.

One approach would be to simply look at the interim implementation conditions for every possible type space $\mathcal{T}$ and then try to characterize the intersection or union of these conditions for all type spaces. This is the approach we initially pursued, and it works in a brute force kind of way. In Section 6.1, we review what happens under this approach.

But we focus our analysis on a second, more elegant, approach. We first establish an equivalence between robust and "rationalizable implementation" and then derive the necessary conditions for robust implementation as an implication of rationalizable implementation. The advantage of the second approach is that after establishing the equivalence, we do not need to argue in terms of large type spaces, but rather derive the results from a novel argument using the iterative elimination process.

Definition 4 (Rationalizable Implementation)
Social choice function f is implemented in rationalizable strategies by mechanism $\mathcal{M}$ if, for all θ, $S^{\mathcal{M}}(\theta) \neq \varnothing$ and if for all θ and m, $m \in S^{\mathcal{M}}(\theta) \Rightarrow g(m) = f(\theta)$.

We now report a formal epistemic argument that relates the rationalizable messages to the set of messages that might be played in any equilibrium on any type space.

Proposition 1 (Rationalizable Actions)
$m_i \in S^{\mathcal{M}}(\theta_i)$ if and only if there exists a type space $\mathcal{T}$, an interim equilibrium σ of the game $(\mathcal{T}, \mathcal{M})$ and a type $t_i \in T_i$ such that (i) $\sigma_i(m_i|t_i) > 0$ and (ii) $\widehat{\theta}_i(t_i) = \theta_i$.

Proof. ($\Rightarrow$) Suppose $m_i^* \in S^{\mathcal{M}}(\theta_i^*)$. Now consider the following type space $\mathcal{T}$ defined through:

$$T_i = \left\{ (m_i, \theta_i) \left| m_i \in S_i^{\mathcal{M}}(\theta_i) \right. \right\}.$$

Let

$$\widehat{\theta}_i(m_i, \theta_i) \triangleq \theta_i.$$

By (2), we know that for each $m_i \in S_i^{\mathcal{M}}(\theta_i)$, there exists $\lambda_i^{m_i,\theta_i} \in \Delta(M_{-i} \times \Theta_{-i})$ such that:

$$\lambda_i^{m_i,\theta_i}(m_{-i}, \theta_{-i}) > 0 \Rightarrow m_j \in S_j^{\mathcal{M}}(\theta_j) \text{ for each } j \neq i;$$

and

$$\sum_{m_{-i},\theta_{-i}} \lambda_i^{m_i,\theta_i}(m_{-i}, \theta_{-i}) \left[u_i(g(m_i, m_{-i}), (\theta_i, \theta_{-i})) - u_i\left(g\left(m_i', m_{-i}\right), (\theta_i, \theta_{-i})\right) \right] \geq 0, \ \forall m_i' \in M_i.$$

Let

$$\widehat{\pi}_i(m_{-i}, \theta_{-i})[m_i, \theta_i] \triangleq \lambda_i^{m_i,\theta_i}(m_{-i}, \theta_{-i}).$$

Now by construction, there is a pure strategy equilibrium s with $s_i(m_i, \theta_i) = m_i$. But now $s_i(m_i^*, \theta_i^*) = m_i^*$ and $\widehat{\theta}(m_i^*, \theta_i^*) = \theta_i^*$.

($\Leftarrow$) Suppose there exists a type space $\mathcal{T}$, an equilibrium σ of $(\mathcal{T}, \mathcal{M})$, and $m_i^* \in M_i$ and $t_i^* \in T_i$ such that $\sigma_i(m_i^* | t_i^*) > 0$ and $\widehat{\theta}_i(t_i^*) = \theta_i^*$. Let

$$S_i(\theta_i) = \left\{ m_i : \sigma_i(m_i | t_i) > 0 \text{ and } \widehat{\theta}_i(t_i) = \theta_i \text{ for some } t_i \in T_i \right\}.$$

Now interim equilibrium conditions ensure that $b(S) \geq S$. Thus $S \leq S^{\mathcal{M}}$. Thus $m_i^* \in S_i^{\mathcal{M}}\left(\widehat{\theta}_i(t_i^*)\right)$, which concludes the proof. ■

Brandenburger and Dekel (1987) showed an equivalence for finite action complete information games between the set of actions surviving iterated deletion of strictly dominant actions and the set of actions that could be played in a subjective correlated equilibrium. Proposition 1 is a straightforward generalization of Brandenburger and Dekel (1987) to incomplete information and infinite actions. The infinite action extension (for complete information) was shown in Lipman (1994). The finite action incomplete information extension is reported in a recent paper of Battigalli and Siniscalchi (2003) (following an earlier analysis in Battigalli (1999)).

3 Finite Mechanisms

A complicating element in using the relationship between equilibrium strategies and rationalizable strategies in the implementation context is the fact that the augmented mechanisms often have

infinite message spaces and that best responses may not exist. These complications are inherent to the entire implementation literature and we therefore have to carefully address these issues before we establish the implementation results. In this section we restrict attention to finite mechanisms, i.e. where each M_i is finite and we extend the argument to infinite mechanisms in the next section. We note that all the results in this section will extend to more general "well-behaved" mechanisms (e.g., compact mechanism or mechanisms where best responses always exist as in Abreu and Matsushima (1992b)).

With finite mechanisms, proposition 1 immediately implies an equivalence between robust and rationalizable implementation.

Corollary 1 (Equivalence)
Social choice function f is robustly implemented by mechanism $\mathcal{M}$ if and only if it is rationalizably implemented by mechanism $\mathcal{M}$.

We now establish necessary conditions for robust implementation which use the equivalence between robust and rationalizable implementation.

3.1 Ex Post Incentive Compatibility

The following ex post incentive compatibility condition is a necessary condition for robust truthful (or partial) implementation as established in Bergemann and Morris (2005b).

Definition 5 (EPIC)
Social choice function f satisfies ex post incentive compatibility (EPIC) if

$$u_i\left(f\left(\theta_i,\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right)\geq u_i\left(f\left(\theta_i',\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right),$$

for all i, θ_i, θ_i' and θ_{-i}.

In the context of robust (full) implementation, we require a strict version of the ex post incentive compatibility conditions.

Definition 6 (Semi-Strict EPIC)
Social choice function f satisfies semi-strict ex post incentive compatibility (semi-strict EPIC) if, for each i, θ_i, θ_i', θ_{-i},

$$u_i\left(f\left(\theta_i,\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right)>u_i\left(f\left(\theta_i',\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right),$$

if there exists $\theta_{-i}'\in\Theta_{-i}$ such that $f\left(\theta_i,\theta_{-i}'\right)\neq f\left(\theta_i',\theta_{-i}'\right)$.

The necessity of the semi-strict version of ex post incentive compatibility now follows directly from the conditions imposed by rationalizable implementation.

Proposition 2 (Necessity of Semi-Strict EPIC)
If social choice function f is robustly implementable by a finite mechanism, then f satisfies semi-strict EPIC.

Proof. If mechanism $\mathcal{M}$ robustly implements f, then, for each i, there exists $m_i^* : \Theta_i \to M_i$ such that

$$g\left(m^*\left(\theta\right)\right) = f\left(\theta\right) \text{ and } m^*\left(\theta\right) \in S^{\mathcal{M}}\left(\theta\right);$$

(we can simply let $m_i^*\left(\theta_i\right)$ be any element of $S_i^{\mathcal{M}}\left(\theta_i\right)$).

Suppose semi-strict EPIC fails. Then there exists i, θ and θ' such that:

$$f\left(\theta'\right) \neq f\left(\theta_i, \theta'_{-i}\right) \tag{3}$$

and

$$u_i\left(f\left(\theta'_i, \theta_{-i}\right), \theta\right) \geq u_i\left(f\left(\theta\right), \theta\right). \tag{4}$$

Now (4) implies that:

$$\begin{aligned} u_i\left(g\left(m_i^*\left(\theta'_i\right), m_{-i}^*\left(\theta_{-i}\right)\right), \left(\theta_i, \theta_{-i}\right)\right) &= u_i\left(f\left(\theta'_i, \theta_{-i}\right), \theta\right) \\ &\geq u_i\left(f\left(\theta\right), \theta\right) \\ &= u_i\left(g\left(m_i^*\left(\theta_i\right), m_{-i}^*\left(\theta_{-i}\right)\right), \left(\theta_i, \theta_{-i}\right)\right). \end{aligned}$$

Since $m_i^*\left(\theta_i\right) \in S_i^{\mathcal{M}}\left(\theta_i\right)$, this implies $m_i^*\left(\theta'_i\right) \in S_i^{\mathcal{M}}\left(\theta_i\right)$. But now

$$f\left(\theta_i, \theta'_{-i}\right) = g\left(m_i^*\left(\theta'_i\right), m_{-i}^*\left(\theta'_{-i}\right)\right) = f\left(\theta'\right),$$

contradicting (3). ■

Next we present two related, yet distinct, monotonicity conditions which are at the core of the robust implementation results.

3.2 Robust Monotonicity

To understand the robust monotonicity condition, it is useful to first think about agents playing the direct mechanism. In the direct mechanism, an agent i may or may not report truthfully. A *deception* is a set-valued profile $\beta = (\beta_1, ..., \beta_I)$, where

$$\beta_i : \Theta_i \to 2^{\Theta_i} / \varnothing,$$

with $\theta_i \in \beta_i(\theta_i)$ for all i and all θ_i. A deception of agent i with payoff type θ_i is a set of possible reports by agent i. By definition, a deception of payoff type θ_i includes, but is not restricted to, θ_i itself.

Definition 7 (Acceptable / Unacceptable Deception)
A deception is acceptable if $\theta' \in \beta(\theta) \Rightarrow f(\theta') = f(\theta)$. *A deception is unacceptable if it is not acceptable.*

In this language, the "truthtelling" deception, defined by $\beta_i^*(\theta_i) \triangleq \theta_i$ for all θ_i is an acceptable deception. Other deceptions of agent i may also be acceptable if the social choice function does not vary with respect to some subset of reports of agent i for all type profiles of the other agents. The inverse mapping of a deception β_i represents the set of true type profiles θ_i which could lead to a report θ_i' and we write

$$\beta_i^{-1}(\theta_i') \triangleq \left\{\theta_i \left| \theta_i' \in \beta_i(\theta_i) \right.\right\}.$$

A "robust monotonicity" condition is key to our main result. In the direct mechanism, where agents other than i report themselves to be types θ_{-i}, agent i can obtain outcomes $f(\theta_i', \theta_{-i})$ for any θ_i'. But once we allow augmented mechanisms, we could conceivably offer agent i a larger set of lotteries if he reports deviant behavior of his opponents. We need to identify, for any given report θ_{-i}, the set of lotteries with the property that whatever agent i's actual type, he would never prefer such an allocation to what he would obtain under the social choice function if other agents were reporting truthfully. Thus:

$$Y_i(\theta_{-i}) \triangleq \left\{y \in Y \left| u_i\left(y, (\theta_i', \theta_{-i})\right) \leq u_i\left(f(\theta_i', \theta_{-i}), (\theta_i', \theta_{-i})\right) \text{ for all } \theta_i' \in \Theta_i \right.\right\}. \tag{5}$$

Henceforth, we refer to the set $Y_i(\theta_{-i})$ as the *reward set* (for agent i).

Suppose now that it was common knowledge that in the direct mechanism, type θ_i of agent i will send a report $\theta_i' \in \beta_i(\theta_i)$. If β is acceptable, we would know that f was being implemented. But if β is unacceptable, we must find a type of some agent who is prepared to report that other agents are misreporting. But for the "whistle-blower" who is going to report that we are in a bad equilibrium, we cannot know what he believes about the types of the other agents, nor can we know what message he expects to hear except that it is a message consistent with the deception. We thus have to allow for all possible beliefs ψ_i of agent i over payoff types $\theta_{-i} \in \Theta_{-i}$ consistent with a report θ_{-i}' from a given deception profile β, or

$$\psi_i \in \Delta\left(\beta_{-i}^{-1}(\theta_{-i}')\right).$$

Finally, the reward that he is offered must not mess up the truth-telling behavior in the good equilibrium. This gives the following condition:

Definition 8 (Dual Robust Monotonicity)
Social choice function f satisfies dual robust monotonicity if, for every unacceptable deception β, there exist i, θ_i, $\theta'_i \in \beta_i(\theta_i)$ such that, for all $\theta'_{-i} \in \Theta_{-i}$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta'_{-i}\right)\right)$, there exists $y \in Y$ such that:

$$\sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta'_{-i}\right)} \psi_i(\theta_{-i})\, u_i(y, (\theta_i, \theta_{-i})) > \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta'_{-i}\right)} \psi_i(\theta_{-i})\, u_i\left(f\left(\theta'_i, \theta'_{-i}\right), (\theta_i, \theta_{-i})\right) \tag{6}$$

and for all $\theta''_i \in \Theta_i$:

$$u_i\left(f\left(\theta''_i, \theta'_{-i}\right), \left(\theta''_i, \theta'_{-i}\right)\right) \geq u_i\left(y, \left(\theta''_i, \theta'_{-i}\right)\right). \tag{7}$$

Social choice function f satisfies dual strict *robust monotonicity if for all θ''_i with $f\left(\theta''_i, \theta'_{-i}\right) \neq y$ the inequality (7) is strict.*

We call this the "dual" version of robust monotonicity because, in the special case where the pure outcome space Z and payoff type spaces Θ_i are finite, dual robust monotonicity can be given a simpler expression.

Definition 9 (Robust Monotonicity)
Social choice function f satisfies robust monotonicity if, for every unacceptable deception β, there exist i, θ_i, $\theta'_i \in \beta_i(\theta_i)$ such that, for all $\theta'_{-i} \in \Theta_{-i}$ there exists y such that for all $\theta_{-i} \in \beta_{-i}^{-1}\left(\theta'_{-i}\right)$:

$$u_i(y, (\theta_i, \theta_{-i})) > u_i\left(f\left(\theta'_i, \theta'_{-i}\right), (\theta_i, \theta_{-i})\right)$$

and for all $\theta''_i \in \Theta_i$:

$$u_i\left(f\left(\theta''_i, \theta'_{-i}\right), \left(\theta''_i, \theta'_{-i}\right)\right) \geq u_i\left(y, \left(\theta''_i, \theta'_{-i}\right)\right). \tag{8}$$

Social choice function f satisfies strict *robust monotonicity if for all θ''_i with $f\left(\theta''_i, \theta'_{-i}\right) \neq y$ the inequality (8) is strict.*

The equivalence of robust monotonicity and dual robust monotonicity when the pure outcome space Z and payoff type spaces Θ_i are finite is established in Lemma 2 in the appendix.

Proposition 3 (Necessity of Dual Strict Robust Monotonicity)
If social choice function f is robustly implementable by a finite mechanism, then f satisfies dual strict robust monotonicity.

Proof. Fix an unacceptable deception β. Let $\widehat{k}$ be the largest k such that for every i, θ_i and $\theta'_i \in \beta_i(\theta_i)$,

$$S_i^{\mathcal{M}}\left(\theta'_i\right) \subseteq S_i^{\mathcal{M},\hat{k}}(\theta_i).$$

We know that such a $\widehat{k}$ exists because $S_i^{\hat{k}} \cap S_i^{\mathcal{M}}\left(\theta_i'\right) = S_i^{\mathcal{M}}\left(\theta_i'\right)$ and, since $\mathcal{M}$ implements f, we must have $S_i^{\mathcal{M}}(\theta_i) \cap S_i^{\mathcal{M}}\left(\theta_i'\right) = \varnothing$. Now we know that there exists i and $\theta_i' \in \beta_i(\theta_i)$ such that

$$S_i^{\mathcal{M},\hat{k}+1}(\theta_i) \cap S_i^{\mathcal{M}}\left(\theta_i'\right) \neq S_i^{\mathcal{M}}\left(\theta_i'\right).$$

Let

$$\widehat{m}_i \in S_i^{\mathcal{M},\hat{k}}(\theta_i) \cap S_i^{\mathcal{M}}\left(\theta_i'\right),$$

and

$$\widehat{m}_i \notin S_i^{\mathcal{M},\hat{k}+1}(\theta_i) \cap S_i^{\mathcal{M}}\left(\theta_i'\right).$$

Since message $\widehat{m}_i$ gets deleted for θ_i at round $\widehat{k}+1$, we know that for every $\lambda_i \in \Delta(M_{-i} \times \Theta_{-i})$ such that

$$\lambda_i(m_{-i}, \theta_{-i}) > 0 \Rightarrow m_j \in S_j^{\mathcal{M},\hat{k}}(\theta_j) \text{ for all } j \neq i,$$

there exists m_i^* such that

$$\sum_{m_{-i},\theta_{-i}} \lambda_i(m_{-i}, \theta_{-i})\, u_i\left(g\left(m_i^*, m_{-i}\right), (\theta_i, \theta_{-i})\right) > \sum_{m_{-i},\theta_{-i}} \lambda_i(m_{-i}, \theta_{-i})\, u_i\left(g\left(\widehat{m}_i, m_{-i}\right), (\theta_i, \theta_{-i})\right).$$

Let

$$\widehat{m}_j \in S_j^{\mathcal{M}}\left(\theta_j'\right),$$

for all $j \neq i$. Now the above claim remains true if we restrict attention to distributions λ_i putting probability 1 on $\widehat{m}_{-i}$. Thus for every $\psi_i \in \Delta(\Theta_{-i})$ such that

$$\psi_i(\theta_{-i}) > 0 \Rightarrow \widehat{m}_j \in S_j^{\mathcal{M},\hat{k}}(\theta_j) \text{ for all } j \neq i,$$

there exists m_i^* such that

$$\sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i\left(g\left(m_i^*, \widehat{m}_{-i}\right), (\theta_i, \theta_{-i})\right) > \sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i\left(g\left(\widehat{m}_i, \widehat{m}_{-i}\right), (\theta_i, \theta_{-i})\right).$$

But $\widehat{m} \in S^{\mathcal{M}}\left(\theta'\right)$, so (since $\mathcal{M}$ robustly implements f), $g\left(\widehat{m}_i, \widehat{m}_{-i}\right) = f\left(\theta'\right)$. Also observe that if $\theta_{-i}' \in \beta_{-i}(\theta_{-i})$, then $\widehat{m}_{-i} \in S_{-i}^{\mathcal{M},\hat{k}}(\theta_{-i})$. Thus for every $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right)$, there exists m_i^* such that

$$\sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i\left(g\left(m_i^*, \widehat{m}_{-i}\right), (\theta_i, \theta_{-i})\right) > \sum_{\theta_{-i}} \psi_i(\theta_{-i})\, u_i\left(f\left(\theta'\right), (\theta_i, \theta_{-i})\right),$$

which establishes the reward inequality, (6), of dual strict robust monotonicity.

Now suppose the incentive inequalities, (7), are not satisfied strictly, and hence:

$$u_i\left(g\left(m_i^*, \widehat{m}_{-i}\right), \left(\widetilde{\theta}_i, \theta_{-i}'\right)\right) \geq u_i\left(f\left(\widetilde{\theta}_i, \theta_{-i}'\right), \left(\widetilde{\theta}_i, \theta_{-i}'\right)\right)$$

and $g\left(m_i^*, \widehat{m}_{-i}\right) \neq f\left(\widetilde{\theta}_i, \theta'_{-i}\right)$. Now, for any

$$m_i \in \arg\max_{m'_i} u_i\left(g\left(m'_i, \widehat{m}_{-i}\right), \left(\widetilde{\theta}_i, \theta'_{-i}\right)\right), \tag{9}$$

since $\widehat{m}_{-i} \in S_i^{\mathcal{M}}\left(\theta'_{-i}\right)$, we must have $m_i \in S_i^{\mathcal{M}}\left(\widetilde{\theta}_i\right)$ and thus $g\left(m_i, \widehat{m}_{-i}\right) = f\left(\theta_i, \theta'_{-i}\right)$. Thus from (9) we also know that m_i^* achieves the maximum:

$$m_i^* \in \arg\max_{m'_i} u_i\left(g\left(m'_i, \widehat{m}_{-i}\right), \left(\widetilde{\theta}_i, \theta'_{-i}\right)\right)$$

and, for all $\widetilde{\theta}_i$, if

$$u_i\left(g\left(m_i^*, \widehat{m}_{-i}\right), \left(\widetilde{\theta}_i, \theta'_{-i}\right)\right) \geq u_i\left(f\left(\widetilde{\theta}_i, \theta'_{-i}\right), \left(\widetilde{\theta}_i, \theta'_{-i}\right)\right),$$

then $g\left(m_i^*, \widehat{m}_{-i}\right) = f\left(\widetilde{\theta}_i, \theta'_{-i}\right)$.

Now setting $y \triangleq g\left(m_i^*, \widehat{m}_{-i}\right)$, we have established that for each $\theta'_{-i} \in \beta_{-i}\left(\theta_{-i}\right)$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta'_{-i}\right)\right)$, there exists y such that $y \in Y_i\left(\theta'_{-i}\right)$ and

$$\sum_{\theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(g\left(m_i^*, \widehat{m}_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right) > \sum_{\theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(f\left(\theta'\right), \left(\theta_i, \theta_{-i}\right)\right),$$

which concludes the proof. ■

3.3 Robust Measurability

We now present a distinct necessary condition for robust implementation. We will be interested in the set of preferences that an agent might have if his payoff type is θ_i and he knows that the type θ_j of each opponent j belongs to some subset Ψ_j of his payoff types Θ_j. Write $\mathcal{R}$ for the set of expected utility preference relations on lotteries Y. We will write $R_{\theta_i,\psi_i} \in \mathcal{R}$ for the preference relation of agent i if his payoff type is θ_i and he has belief $\psi_i \in \Delta\left(\Theta_{-i}\right)$ about the types of others:

$$\forall y, y' \in Y: \qquad y R_{\theta_i,\psi_i} y' \quad \Leftrightarrow \quad \sum_{\theta_{-i} \in \Theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(y, \left(\theta_i, \theta_{-i}\right)\right) \geq \sum_{\theta_{-i} \in \Theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(y', \left(\theta_i, \theta_{-i}\right)\right);$$

We write $\mathcal{R}_i\left(\theta_i, \Psi_{-i}\right)$ for the set of preferences agent i might have if his payoff type is θ_i and he might have any beliefs over others' payoff types.

$$\mathcal{R}_i\left(\theta_i, \Psi_{-i}\right) = \left\{R \in \mathcal{R} \,\middle|\, R = R_{\theta_i,\psi_i} \text{ for some } \psi_i \in \Delta\left(\Psi_{-i}\right)\right\}.$$

Say that type set profile Ψ_{-i} *separates* Ψ_i if

$$\bigcap_{\theta_i \in \Psi_i} \mathcal{R}_i\left(\theta_i, \Psi_{-i}\right) = \varnothing.$$

Let $\Xi = \left(\Xi_i\right)_{i=1}^I \in \times_{i=1}^I 2^{\Theta_i}$ be a profile of type sets for each agent. Say that Ξ is *mutually inseparable* if, for each i and $\Psi_i \in \Xi_i$, there exists $\Psi_{-i} \in \Xi_{-i}$ such that Ψ_{-i} does not separate Ψ_i.

Definition 10 (Robust Measurability)
Social choice function f satisfies robust measurability if Ξ mutually inseparable, $\Psi_i \in \Xi_i$ and $\left\{\theta_i', \theta_i''\right\} \subseteq \Psi_i \Rightarrow f\left(\theta_i', \theta_{-i}\right) = f\left(\theta_i'', \theta_{-i}\right)$ for all θ_{-i}.

If payoff types are finite, one can give an alternative iterative definition of robust measurability: let $\Xi_i^0 = 2^{\Theta_i}$,

$$\Xi_i^{k+1} = \left\{\Psi_i \in \Xi_i^k \mid \Psi_{-i} \text{ does not separate } \Psi_i, \text{ for some } \Psi_{-i} \in \Xi_{-i}^k\right\}, \tag{10}$$

and

$$\Xi_i^* = \bigcap_{k \geq 0} \Xi_i^k; \tag{11}$$

now social choice function f satisfies robust measurability if $\left\{\theta_i', \theta_i''\right\} \in \Xi_i^* \Rightarrow f\left(\theta_i', \theta_{-i}\right) = f\left(\theta_i'', \theta_{-i}\right)$ for all θ_{-i}.[9]

Proposition 4 (Necessity of Robust Measurability)
If social choice function f is robustly implementable by a finite mechanism, then f satisfies robust measurability.

Proof. Since f is robustly implementable, there exists a mechanism $\mathcal{M}$ such that

$$m \in S^{\mathcal{M}}(\theta) \Rightarrow g(m) = f(\theta).$$

Now suppose Ξ is mutually inseparable. We argue by induction that, for all i, $\Psi_i \in \Xi_i$ and k there exists a set of messages $\varnothing \neq M_i^k(\Psi_i) \subseteq S_i^{\mathcal{M},k}(\theta_i)$ for all $\theta_i \in \Psi_i$. This is true by definition for $k = 0$. Suppose that it is true for k. Now Ξ mutually inseparable implies that for any $\Psi_i \in \Xi_i$, there exists $\Psi_{-i} \in \Xi_{-i}$, R and, for each $\theta_i \in \Psi_i$, $\lambda_i^{\theta_i} \in \Delta(\Psi_{-i})$ such that $R_{\theta_i, \lambda_i^{\theta_i}} = R$. Now let $M_i^{k+1}(\Psi_i)$ be the optimal messages of agent i when he believes that his opponents will sent some message profile in $M_{-i}^k(\Psi_{-i})$ with probability 1 and has beliefs $\lambda_i^{\theta_i}$ about the type profile of his opponents, i.e.,

$$M_i^{k+1}(\Psi_i) = \bigcup_{m_{-i} \in M_{-i}^k(\Psi_{-i})} \arg\max_{m_i'} \sum_{\theta_{-i}} \lambda_i^{\theta_i}(\theta_{-i}) u_i\left(g\left(m_i', m_{-i}\right), (\theta_i, \theta_{-i})\right).$$

By construction, $\varnothing \neq M_i^{k+1}(\Psi_i) \subseteq S_i^{\mathcal{M},k}(\theta_i)$ for all $\theta_i \in \Psi_i$. Now for each $\Psi_i \in \Xi_i$, $M_i^k(\Psi_i)$ is a decreasing sequence under set inclusion. Since M_i is finite, there exists $\varnothing \neq M_i^*(\Psi_i) = \bigcap_{k \geq 0} M_i^k(\Psi_i)$. Thus $M_i^*(\Psi_i) \subseteq S_i^{\mathcal{M}}(\theta_i)$ for all $\theta_i \in \Psi_i$. Now if $\left\{\theta_i', \theta_i''\right\} \subseteq \Psi_i$, there exists $m_i \in M_i^*(\Psi_i) \subseteq S_i^{\mathcal{M}}\left(\theta_i'\right)$ and $m_i \in M_i^*(\Psi_i) \subseteq S_i^{\mathcal{M}}\left(\theta_i''\right)$. Now fix any $m_{-i} \in S_{-i}^{\mathcal{M}}(\theta_{-i})$, and

[9] See Lemma 3 of Bergemann and Morris (2008c).

we have $(m_i, m_{-i}) \in S^{\mathcal{M}}\left(\theta_i', \theta_{-i}\right) \Rightarrow g\left(m_i, m_{-i}\right) = f\left(\theta_i', \theta_{-i}\right)$ and $(m_i, m_{-i}) \in S^{\mathcal{M}}\left(\theta_i'', \theta_{-i}\right) \Rightarrow g\left(m_i, m_{-i}\right) = f\left(\theta_i'', \theta_{-i}\right)$. Thus $f\left(\theta_i', \theta_{-i}\right) = f\left(\theta_i'', \theta_{-i}\right)$. ■

In Appendix B, we show by means of two examples that robust monotonicity does not imply nor is it implied by robust measurability.

We have pursued two ways of deriving sufficient conditions in prior work. First, we identified natural restrictions on the environment that make these necessary conditions sufficient (Bergemann and Morris (2007)). Second, we showed what happened if we weaken the implementation requirement to virtual implementation (Bergemann and Morris (2008c)). We briefly review these results below. If we neither put restrictions on the environment nor allow virtual implementation, then we do not know how to derive tight sufficient conditions for finite, or other well-behaved, mechanisms. However, as in the existing complete information and standard Bayesian implementation literature, it is possible to obtain tight conditions if we allow for badly behaved mechanisms. These results are reported in the remainder of the paper. We believe they improve our understanding about how the different elements in the incomplete information implementation literature fit together and highlight the role of infinite mechanisms.

3.4 Single Crossing Aggregator Environments

In Bergemann and Morris (2007), we consider payoff environments in which each payoff type space Θ_i is completely ordered and where there exist for each i, an aggregator function $h_i : \Theta \to \mathbb{R}$ and a valuation function $v_i : Y \times \mathbb{R} \to \mathbb{R}$ such that

$$v_i\left(y, h_i\left(\theta\right)\right) \triangleq u_i\left(y, \theta\right), \tag{12}$$

where h_i is continuous and strictly increasing in θ_i and $v_i : Y \times \mathbb{R} \to \mathbb{R}$ is continuous and satisfies the following strict single crossing property: for all $\phi < \phi' < \phi''$,

$$v_i\left(y, \phi\right) > v_i\left(y', \phi\right) \text{ and } v_i\left(y, \phi'\right) = v_i\left(y', \phi'\right) \quad \Rightarrow \quad v_i\left(y, \phi''\right) < v_i\left(y', \phi''\right). \tag{13}$$

The aggregator functions $h = \left(h_i\right)_{i=1}^I$ are said to satisfy the *contraction property* if, for all deceptions $\beta \neq \beta^*$, there exists i, θ_i and $\theta_i' \in \beta_i\left(\theta_i\right)$ with $\theta_i' \neq \theta_i$, such that

$$\text{sign}\left(\theta_i - \theta_i'\right) = \text{sign}\left(h_i\left(\theta_i, \theta_{-i}\right) - h_i\left(\theta_i', \theta_{-i}'\right)\right) \tag{14}$$

for all θ_{-i} and $\theta_{-i}' \in \beta_{-i}\left(\theta_{-i}\right)$. In single crossing aggregator environments as described by (12) and (13), the contraction property is equivalent to both dual strict robust monotonicity and robust measurability.

We say that a social choice function f is *responsive* if for all $\theta_i \neq \theta_i'$, there exists θ_{-i} such that $f\left(\theta_i, \theta_{-i}\right) \neq f\left(\theta_i', \theta_{-i}\right)$. If a social choice function is responsive, then semi-strict EPIC simplifies

to *strict ex post incentive compatibility*, i.e., $u_i\left(f\left(\theta_i,\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right) > u_i\left(f\left(\theta_i',\theta_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right)$, for all i, $\theta_i \neq \theta_i'$ and θ_{-i}.

Proposition 5 (Contraction Property)
In a single crossing aggregator environment, a responsive social choice function f is robustly implementable if and only if it satisfies strict ex post incentive compatibility and the contraction property.

This result is reported in Theorem 1 and 2 of Bergemann and Morris (2007). It follows that the necessary conditions of propositions 2, 3 and 4 are also sufficient in these environments. Note that in the discrete type setting of this paper, the continuity properties are automatically satisfied if the payoff type spaces are finite. Bergemann and Morris (2007) allowed for compact payoff type spaces and pure outcome spaces. Bergemann and Morris (2007) also showed that when robust implementation is possible, it is possible in a "direct" mechanism where agents report just their payoff types.

3.5 Robust Virtual Implementation

The necessary conditions for robust implementation also become sufficient conditions if we relax the requirement from (robust) exact to (robust) virtual implementation. In Bergemann and Morris (2008c), we consider settings where the space of pure outcomes and payoff types are finite. By Corollary 1 we can therefore define robust virtual implementation directly with reference to the rationalizable messages in a given mechanism $\mathcal{M}$.

Definition 11 (Robust Virtual Implementation)
Social choice function f is robustly virtually implementable if, for each $\varepsilon > 0$, there exists a mechanism $\mathcal{M}$ such that for all θ, $S^{\mathcal{M}}(\theta) \neq \varnothing$ and if for all θ and m, $m \in S^{\mathcal{M}}(\theta) \Rightarrow \|g(m) - f(\theta)\| \leq \varepsilon$.

We established in Theorem 1 and 2 of Bergemann and Morris (2008c) the following necessary and sufficient conditions for robust virtual implementation.

Proposition 6 (Robust Measurability)
A social choice function is robustly virtually implementable if and only if it satisfies ex post incentive compatibility and robust measurability.

Thus strict robust monotonicity can be dropped and semi-strict EPIC can be weakened to EPIC if virtual implementation is enough. With these weakenings, the necessary conditions are sufficient.

3.6 A Coordination Example

We conclude this section with an example that demonstrates that while robust implementation is a strong requirement, it is weaker than dominant strategies. In the example there are two agents, $i = 1, 2$. Each agent i has two possible types, θ_i and θ_i'. There are six possible outcomes: $Z = \{a, b, c, d, z, z'\}$. The payoffs of the agents are a function of the allocation and the true payoff type profile, given by:

a	θ_2	θ_2'
θ_1	$3,3$	$0,0$
θ_1'	$0,0$	$1,1$

b	θ_2	θ_2'
θ_1	$0,0$	$3,3$
θ_1'	$1,1$	$0,0$

c	θ_2	θ_2'
θ_1	$0,0$	$1,1$
θ_1'	$3,3$	$0,0$

d	θ_2	θ_2'
θ_1	$1,1$	$0,0$
θ_1'	$0,0$	$3,3$

(15)

and

z	θ_2	θ_2'
θ_1	$2,2$	$2,0$
θ_1'	$2,2$	$2,0$

z′	θ_2	θ_2'
θ_1	$2,0$	$2,2$
θ_1'	$2,0$	$2,2$

.

The social choice function is given by the efficient outcome at each type profile:

f	θ_2	θ_2'
θ_1	a	b
θ_1'	c	d

. (16)

Clearly, the social choice function is strictly ex post incentive compatible. But in the "direct mechanism" where each agent simply reports his type, there will always be an equilibrium where each type of each agent misreports his type, and each agent gets a payoff of 1. This is also strictly ex post incentive compatible. The social choice function f which selects among $\{a, b, c, d\}$ embeds a coordination game. We further observe that the payoff for agent 1 from allocations z and z' are equal and constant for all type profiles. On the other hand, the payoff of agent 2 from z and z' depends on his type but not on the type of the other agent.

We now consider the following augmented, but finite, mechanism which responds to the messages of the agents as follows:

g	θ_2	θ_2'
θ_1	a	b
θ_1'	c	d
ζ	z	z'

The augmented mechanism enriches the message space of agent 1 by a single message ζ. The

corresponding incomplete information game has the following payoffs:

	type	θ_2		θ_2'	
type	action	θ_2	θ_2'	θ_2	θ_2'
θ_1	θ_1	$3,3$	$0,0$	$0,0$	$3,3$
	θ_1'	$0,0$	$1,1$	$1,1$	$0,0$
	ζ	$2,2$	$2,0$	$2,0$	$2,2$
θ_1'	θ_1	$0,0$	$1,1$	$1,1$	$0,0$
	θ_1'	$3,3$	$0,0$	$0,0$	$3,3$
	ζ	$2,2$	$2,0$	$2,0$	$2,2$

Suppose we iteratively remove actions for each type that could never be a best response given the type action profiles remaining. Thus in the first round, we would observe that type θ_1 would never send message θ_1' and type θ_1' would never send message θ_1. Knowing this, we could conclude that type θ_2 would never send message θ_2' and type θ_2' would never send message θ_2. This in turn implies that neither type of agent 1 will ever send message ζ. Thus the only remaining message for each type of each agent is truth-telling. But now they must behave this way in any equilibrium on any type space.

4 Rationalizable and Robust Implementation in Infinite Mechanisms

In Section 3 we established the equivalence between rationalizable and robust implementation for finite mechanisms. A complicating factor is that augmented mechanisms often have infinite message spaces and so best responses may not exist. We now address these issues for infinite mechanisms and then establish the implementation results for general mechanisms.

4.1 Best Response

We observe that with infinite mechanisms there is no a priori guarantee that $S^{\mathcal{M}}(\theta_i)$ is non-empty or that a game of incomplete information defined by $(\mathcal{T}, \mathcal{M})$ has an interim equilibrium. The epistemic result of Proposition 1 which related the rationalizable messages with the equilibrium messages for some type space continues to hold, vacuously, in these cases. But for implementation results, we care about existence. We introduce the following two conditions that relate existence of equilibrium on all type spaces to the actions surviving iterated deletion. These conditions use the notion of message correspondence S defined in Section 2.4.

Definition 12 (Ex Post Best Response)
Message correspondence S satisfies the ex post best response property if, for all i and $\theta_i \in \Theta_i$, there exists $m_i^* \in S_i(\theta_i)$ such that

$$m_i^* \in \arg\max_{m_i \in M_i} u_i\left(g\left(m_i, m_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right),$$

for all θ_{-i} and $m_{-i} \in S_{-i}(\theta_{-i})$.

We observe that for S to satisfy the ex post best response property, $S_i(\theta_i)$ must be non-empty for all i and all θ_i.

Definition 13 (Interim Best Response)
Message correspondence S satisfies the interim best response property if, for all i and $\psi_i \in \Delta(\Theta_{-i})$, there exists $\lambda_i \in \Delta(M_{-i} \times \Theta_{-i})$ such that:

1. *$\lambda_i(m_{-i}, \theta_{-i}) > 0 \Rightarrow m_j \in S_j(\theta_j)$ for each $j \neq i$;*
2. *for all $\theta_{-i} \in \Theta_{-i}$:*
$$\sum_{m_{-i} \in M_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) = \psi_i\left(\theta_{-i}\right);$$
3. *for all $\theta_i \in \Theta_i$ there exists $m_i^* \in S_i(\theta_i)$ such that*
$$m_i^* \in \arg\max_{m_i \in M_i} \sum_{m_{-i}, \theta_{-i}} \lambda_i\left(m_{-i}, \theta_{-i}\right) u_i\left(g\left(m_i, m_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right).$$

The interim best response property only requires that for every conjecture over payoff type spaces, there exists some beliefs over messages consistent with the message correspondence S, such that a best response is in the message correspondence. In particular, it does not require that a best response exists for all possible beliefs over message profiles. Note that the ex post best response property is a stronger requirement than the interim best response property, but that the interim best response property also implies that $S_i^{\mathcal{M}}(\theta_i)$ is non-empty for all i and θ_i.

Proposition 1 linked every action profile in the set of rationalizable actions to an equilibrium action for some type space $\mathcal{T}$. Proposition 7 strengthens the relationship between rationalizable and equilibrium actions, after imposing some structure on the best response property of rationalizable and equilibrium actions, respectively.

Proposition 7 (Best Response Properties)

1. *If $S^{\mathcal{M}}$ has the ex post best response property, then $(\mathcal{T}, \mathcal{M})$ has an equilibrium for each $\mathcal{T}$.*

2. If $(\mathcal{T},\mathcal{M})$ has an equilibrium for each $\mathcal{T}$, then $S^{\mathcal{M}}$ satisfies the interim best response property.

Proof. (1.) By the ex post best response property, there exists, for each i, $s_i^*:\Theta_i\to M_i$ such that

$$s_i^*(\theta_i)\in \arg\max_{m_i\in M_i} u_i\left(g\left(m_i,s_{-i}^*(\theta_{-i})\right),(\theta_i,\theta_{-i})\right)$$

for all θ_{-i}. Now fix any type space. The strategy profile s with

$$s_i(t_i)=s_i^*\left(\widehat{\theta}_i(t_i)\right)$$

is an equilibrium of the game $(\mathcal{T},\mathcal{M})$.

(2.) Suppose $(\mathcal{T},\mathcal{M})$ has an equilibrium for each $\mathcal{T}$. Fix any i and $\psi_i\in\Delta(\Theta_{-i})$. Fix any type space $\mathcal{T}$ with, for each $\theta_i\in\Theta_i$, a type $t_i^*(\theta_i)$ such that (a) $\widehat{\theta}_i(t_i^*(\theta_i))=\theta_i$ for each θ_i, (b) there exists $\pi_i\in\Delta(T_{-i})$ such that $\widehat{\pi}_i(t_i^*(\theta_i))=\pi_i$ for all θ_i and (c)

$$\sum_{\left\{t_{-i}:\widehat{\theta}_{-i}(t_{-i})=\theta_{-i}\right\}}\pi_i(t_{-i})=\psi_i(\theta_{-i}) \tag{17}$$

for all θ_i and θ_{-i}. The game has an equilibrium σ. Let m_i be any message with $\sigma_i(m_i|t_i^*(\theta_i))>0$. Let

$$\lambda_i(m_{-i},\theta_{-i})=\sum_{\left\{t_{-i}\in T_{-i}:\widehat{\theta}_{-i}(t_{-i})=\theta_{-i}\right\}}\sigma_{-i}(m_{-i}|t_{-i})\,\pi_i(t_{-i}).$$

Now $\sigma_i(m_i|t_i^*(\theta_i))>0$ implies

$$m_i(\theta_i)\in\arg\max_{m_i\in M_i}\sum_{m_{-i},\theta_{-i}}\lambda_i(m_{-i},\theta_{-i})\,u_i(g(m_i,m_{-i}),(\theta_i,\theta_{-i})).$$

Proposition 1 implies that every message profile m_j which is played in equilibrium by type θ_j is part of the set $S^{\mathcal{M}}$, or that:

$$\lambda_i(m_{-i},\theta_{-i})>0\Rightarrow m_j\in S_j^{\mathcal{M}}(\theta_j)\ \text{ for each } j\neq i.$$

By construction of the type space $\mathcal{T}$, in particular property (c) as expressed by (17), this implies that

$$\sum_{m_{-i}\in M_{-i}}\lambda_i(m_{-i},\theta_{-i})=\psi_i(\theta_{-i})\ \text{ for all } \theta_{-i}\in\Theta_{-i}.$$

Since these properties hold for arbitrary i and $\psi_i\in\Delta(\Theta_{-i})$, $S^{\mathcal{M}}$ satisfies the interim best response property, which concludes the proof. ■

It is unfortunate that there is a gap between the necessary and sufficient conditions in the above proposition. However, an example in the appendix shows that it is possible to construct (admittedly silly) mechanisms where $(\mathcal{T},\mathcal{M})$ has an equilibrium for each $\mathcal{T}$, but $S^{\mathcal{M}}$ fails the ex post best response property.

4.2 Material Implementation

We can maintain the relationship between rationalizable and robust implementation, despite the possibility of non-existence of an interim best response, by qualifying the implementation condition as being "material".

Definition 2M (Material Interim Implementation).
Social choice function f is materially interim implemented on type space $\mathcal{T}$ by mechanism $\mathcal{M}$ if every equilibrium σ of the game $(\mathcal{T}, \mathcal{M})$ satisfies

$$\sigma(m|t) > 0 \Rightarrow g(m) = f\left(\widehat{\theta}(t)\right),$$

for all t.

In contrast to the earlier definition of interim implementation, given in Definition 2, we allow the premise of the definition to be vacuous. In other words, the mechanism $\mathcal{M}$ might have the property that on a given type space, there is no equilibrium. Our terminology mirrors the language of modal logic where proposition A materially implies B whenever A is false, as well as when both A and B are true, see Hughes and Creswell (1996). We similarly weaken the definition of robust and rationalizable implementation.

Definition 3M (Material Robust Implementation).
Social choice function f is materially robustly implemented by mechanism $\mathcal{M}$ if, for every $\mathcal{T}$, f is materially interim implemented on type space $\mathcal{T}$ by mechanism $\mathcal{M}$.

Definition 4M (Material Rationalizable Implementation).
Social choice function f is materially rationalizably implemented by mechanism $\mathcal{M}$ if for all θ and m, $m \in S^{\mathcal{M}}(\theta) \Rightarrow g(m) = f(\theta)$.

With these weaker notions of material implementation Proposition 1 now immediately implies an equivalence between material robust and material rationalizable implementation in the presence of infinite mechanisms.

Corollary 2 (Equivalence)
Social choice function f is materially rationalizably implemented by $\mathcal{M}$ if and only if f is materially robustly implemented by mechanism $\mathcal{M}$.

Proposition 7 gave the slightly messier result relating equilibrium existence and properties of messages surviving iterated deletion. The following corollary gives the immediate implications for our implementation definitions:

Corollary 3 (Necessary Conditions)

1. *If social choice function f is materially rationalizably implemented by mechanism $\mathcal{M}$ and $S^{\mathcal{M}}$ satisfies the ex post best response property, then f is robustly implemented by $\mathcal{M}$.*

2. *If f is robustly implemented by $\mathcal{M}$, then f is materially rationalizably implemented by mechanism $\mathcal{M}$ and $S^{\mathcal{M}}$ satisfies the interim best response property.*

The 'material' qualification will only be used in the necessity part of Theorem 1 where we shall invoke the second part of Corollary 3. There we shall use the fixed-point property of $S^{\mathcal{M}}$, stated earlier in (2), to derive the robust monotonicity condition. In the sufficiency part of the proof, a non-empty set $S^{\mathcal{M}}$ is obtained in the augmented mechanism by virtue of ex post incentive compatibility. The following implication of rationalizable implementation will be used to establish robust monotonicity in Theorem 1.

Lemma 1 (Truthtelling as Best Response)
If f is materially rationalizably implemented by mechanism $\mathcal{M}$ and $S^{\mathcal{M}}$ satisfies the interim best response property, then for all i and $\theta_{-i} \in \Theta_{-i}$, there exists $\nu_i \in \Delta\left(S_{-i}^{\mathcal{M}}(\theta_{-i})\right)$ such that

$$u_i\left(f\left(\theta_i, \theta_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right) \geq \sum_{m_{-i}} \nu_i\left(m_{-i}\right) u_i\left(g\left(m_i, m_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right) \tag{18}$$

for all $m_i \in M_i$ and $\theta_i \in \Theta_i$.

Proof. Applying the definition of the interim best response property for i and the degenerate distribution putting probability 1 on θ_{-i}, we have that there exists $\nu_i \in \Delta\left(S_{-i}^{\mathcal{M}}(\theta_{-i})\right)$ such that

$$\varnothing \neq \arg\max_{m_i} \sum_{m_{-i}, \theta_{-i}} \nu_i\left(m_{-i}\right) u_i\left(\left(m_i, m_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right) \subseteq S_i^{\mathcal{M}}\left(\theta_i\right) \text{ for all } \theta_i \in \Theta_i.$$

But by material rationalizable implementability, $m \in S^{\mathcal{M}}(\theta) \Rightarrow g(m) = f(\theta)$. So

$$u_i\left(f\left(\theta_i, \theta_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right) \geq \sum_{m_{-i}} \nu_i\left(m_{-i}\right) u_i\left(\left(m_i, m_{-i}\right), \left(\theta_i, \theta_{-i}\right)\right),$$

for all $m_i \in M_i$ and $\theta_i \in \Theta_i$. ■

Lemma 1 shows how small the gap between the ex post and interim best response property is. It establishes that truthtelling is a best response against some beliefs over messages m_{-i} for any given payoff type profile θ_{-i}.

5 Infinite Mechanisms

We will need a very weak economic condition to ensure that it is always possible to reward and punish each agent independently of the other agents.

Definition 14 (Conditional No Total Indifference)
The conditional no total indifference (NTI) property is satisfied if, for all i, all θ and all $\psi_i \in \Delta(\Theta_{-i})$, there exists $y, y' \in Y_i(\theta_{-i})$ such that

$$\sum_{\theta'_{-i} \in \Theta_{-i}} \psi_i(\theta_{-i}) u_i\left(y, \left(\theta_i, \theta'_{-i}\right)\right) > \sum_{\theta'_{-i} \in \Theta_{-i}} \psi_i(\theta_{-i}) u_i\left(y', \left(\theta_i, \theta'_{-i}\right)\right).$$

The *conditional no total indifference property* imposes a very weak restriction on the preferences. For example, if there are a finite number of pure outcomes and an agent is never completely indifferent between all lotteries, then we can always find interior outcomes y and y' such that the conditional no total indifference condition is met. The conditional NTI property, together with the use of lotteries, renders an additional no veto property, which typically appears in the sufficient conditions, obsolete. In addition, we can omit the usual cardinality assumption of $I \geq 3$. A related no total indifference condition appears in the context of virtual implementation in Duggan (1997), who requires it to hold at every ex post profile θ and in Serrano and Vohra (2005), who require it at the interim level for a given belief $\psi_i(\theta_{-i})$ of player i.

Theorem 1 (Robust Implementation)

1. *If f is robustly implementable, then f satisfies EPIC and dual robust monotonicity;*
2. *If f satisfies EPIC, dual robust monotonicity and the conditional NTI property, then f is robustly implementable.*

Proof. (1.) We first prove that robust implementability implies EPIC and dual robust monotonicity. We do so by appealing to the necessary conditions for robust implementation in Corollary 3.

We first establish EPIC. By Lemma 1, for all $m_i \in M_i$ and $\theta_i \in \Theta_i$, there exists $\nu_i \in \Delta\left(S_{-i}^{\mathcal{M}}(\theta_{-i})\right)$,

$$u_i\left(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\right) \geq \sum_{m_{-i}} \nu_i(m_{-i}) u_i\left(g(m_i, m_{-i}), (\theta_i, \theta_{-i})\right),$$

If we choose $m_i \in S_i^{\mathcal{M}}\left(\theta'_i\right)$, material rationalizable implementation implies that $g(m_i, m_{-i}) = f\left(\theta'_i, \theta_{-i}\right)$ for all $m_{-i} \in S_{-i}^{\mathcal{M}}(\theta_{-i})$. So

$$u_i\left(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\right) \geq u_i\left(f\left(\theta'_i, \theta_{-i}\right), (\theta_i, \theta_{-i})\right),$$

for all $\theta_i' \in \Theta_i$.

We next establish dual robust monotonicity. Fix an unacceptable deception β and suppose that f is materially rationalizably implementable. There must exist a message correspondence profile S such that

$$b(S) \leq S,$$

and

$$S_i^{\mathcal{M}}\left(\theta_i'\right) \subseteq S_i(\theta_i), \tag{19}$$

for all i, θ_i and $\theta_i' \in \beta_i(\theta_i)$; but

$$S_i^{\mathcal{M}}\left(\theta_i'\right) \nsubseteq b_i(S)[\theta_i], \tag{20}$$

for all i, θ_i and $\theta_i' \in \beta_i(\theta_i)$. The existence of such an S can be established constructively. Clearly $\overline{S}$ satisfies (19). Iteratively apply the operator b. By rationalizable implementation, there exists k (perhaps transfinite) such that:

$$S \triangleq b^k\left(\overline{S}\right) \tag{21}$$

satisfies (20). Thus there exists k such that $b^k\left(\overline{S}\right)$ satisfies (19) and $b^{k+1}\left(\overline{S}\right)$ satisfies (20).

By (20), simply pick

$$\widehat{m}_i \in S_i(\theta_i) \cap S_i^{\mathcal{M}}\left(\theta_i'\right) \quad \text{and} \quad \widehat{m}_i \notin b_i(S)[\theta_i] \cap S_i^{\mathcal{M}}\left(\theta_i'\right).$$

Since message $\widehat{m}_i \notin b_i(S)[\theta_i]$, we know that for every $\lambda_i \in \Delta(M_{-i} \times \Theta_{-i})$ such that

$$\lambda_i(m_{-i}, \theta_{-i}) > 0 \Rightarrow m_j \in S_j(\theta_j) \text{ for all } j \neq i,$$

there exists m_i^* such that

$$\sum_{m_{-i},\theta_{-i}} \lambda_i(m_{-i}, \theta_{-i})\, u_i\left(g\left(m_i^*, m_{-i}\right), (\theta_i, \theta_{-i})\right) > \sum_{m_{-i},\theta_{-i}} \lambda_i(m_{-i}, \theta_{-i})\, u_i\left(g\left(\widehat{m}_i, m_{-i}\right), (\theta_i, \theta_{-i})\right). \tag{22}$$

Next we identify a particular belief $\lambda_i(m_{-i}, \theta_{-i})$ for which the inequality (22) holds. By (18) in Lemma 1, there exists $\nu_i \in \Delta\left(S_{-i}^{\mathcal{M}}\left(\theta_{-i}'\right)\right)$ such that

$$\sum_{m_{-i}} \nu_i(m_{-i})\, u_i\left(g(m_i, m_{-i}), \left(\theta_i'', \theta_{-i}'\right)\right) \leq u_i\left(f\left(\theta_i'', \theta_{-i}'\right), \left(\theta_i'', \theta_{-i}'\right)\right), \tag{23}$$

for all $m_i \in M_i$ and $\theta_i'' \in \Theta_i$. Thus for any $\psi_i \in \Delta(\Theta_{-i})$, we can set

$$\lambda_i(m_{-i}, \theta_{-i}) = \nu_i(m_{-i})\, \psi_i(\theta_{-i}).$$

Applying the above claim (22), there exists m_i^* such that:

$$\sum_{\theta_{-i},m_{-i}} \psi_i(\theta_{-i})\, \nu_i(m_{-i})\, u_i\left(g\left(m_i^*, m_{-i}\right), (\theta_i, \theta_{-i})\right) > \sum_{\theta_{-i},m_{-i}} \psi_i(\theta_{-i})\, \nu_i(m_{-i})\, u_i\left(g\left(\widehat{m}_i, m_{-i}\right), (\theta_i, \theta_{-i})\right).$$

But $\nu_i(m_{-i}) > 0 \Rightarrow (\widehat{m}_i, m_{-i}) \in S^{\mathcal{M}}(\theta')$, so by material rationalizable implementation:

$$g(\widehat{m}_i, m_{-i}) = f(\theta').$$

We also observe that as we defined S to be the set obtained after the k-th iteration of the operator b, see (21), if $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$, then $\nu_i(m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}(\theta_{-i})$. Thus for every $\psi_i \in \Delta\left(\beta_{-i}^{-1}(\theta'_{-i})\right)$, there exists m_i^* such that

$$\sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}) \nu_i(m_{-i}) u_i(g(m_i^*, m_{-i}), (\theta_i, \theta_{-i})) > \sum_{\theta_{-i}, m_{-i}} \psi_i(\theta_{-i}) \nu_i(m_{-i}) u_i\left(f(\theta'), (\theta_i, \theta_{-i})\right). \tag{24}$$

Now, the inequality (24) essentially establishes guarantees the reward inequality for robust monotonicity. We can complete the argument by letting y be the lottery with

$$y \triangleq \sum_{m_{-i}} g(m_i^*, m_{-i}) \nu_i(m_{-i}).$$

We now have established that for each $\theta'_{-i} \in \beta_{-i}(\theta_{-i})$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}(\theta'_{-i})\right)$, there exists y such that (by (23))

$$u_i\left(y, (\theta''_i, \theta'_{-i})\right) \leq u_i\left(f(\theta''_i, \theta'_{-i}), (\theta''_i, \theta'_{-i})\right),$$

for all $\theta''_i \in \Theta_i$, and thus $y \in Y_i(\theta'_{-i})$.[10] And by (24) we then have:

$$\sum_{\theta_{-i}} \psi_i(\theta_{-i}) u_i(y, (\theta_i, \theta_{-i})) > \sum_{\theta_{-i}} \psi_i(\theta_{-i}) u_i\left(f(\theta'), (\theta_i, \theta_{-i})\right).$$

(2.) We now prove that EPIC, dual robust monotonicity and the conditional NTI property imply robust implementation. We do so by explicitly constructing the implementing mechanism. The mechanism will use "interior" lotteries over the deterministic outcome set Z and over the reward sets $Y_i(\theta_{-i})$. Given an arbitrary labelling of the outcome set $Z = \{z_0, z_1, ..., z_k, ...\}$, we define an "interior" lottery over the set Z by

$$\bar{y} = (\bar{y}_0, \bar{y}_1, ..., \bar{y}_k, ...), \tag{25}$$

where

$$\bar{y}_k \triangleq \Pr(z = z_k) = \frac{\delta^k}{1-\delta},$$

for some $\delta \in (0,1)$. For every given profile θ_{-i}, the reward set $Y_i(\theta_{-i})$ is by construction a convex set with at most a countable number of extreme points. We denote the set of extreme

[10] Note that this step implies that even if we had restricted attention to mechanisms with deterministic outcomes, our robust monotonicity condition would only have established that there exists a lottery (not necessarily a deterministic outcome) sufficient to reward a whistle-blower.

points of $Y_i(\theta_{-i})$ by $Y_i^*(\theta_{-i})$ and for some labelling of the points in the set we have $Y_i^*(\theta_{-i}) = \left\{y_{0,\theta_{-i}}, y_{1,\theta_{-i}}, ..., y_{l,\theta_{-i}}, ...\right\}$. An extreme point $y_{l,\theta_{-i}}$ in $Y_i^*(\theta_{-i})$ may be a deterministic or a random outcome and assigns probability $y_{l,\theta_{-i}}(z_k)$ to the pure outcome z_k. For every reward set $Y_{-i}(\theta_{-i})$, we define a "interior" lottery:

$$\bar{y}_{\theta_{-i}} = \left(\bar{y}_{0,\theta_{-i}}, \bar{y}_{1,\theta_{-i}}, ..., \bar{y}_{k,\theta_{-i}}\right) \tag{26}$$

with

$$\bar{y}_{k,\theta_{-i}} \triangleq \frac{1}{1-\delta}\sum_{l=0}^{\infty}\delta^l y_{l,\theta_{-i}}(z_k),$$

where the lottery $\bar{y}_{\theta_{-i}}$ is a compound lottery.

Each agent i sends a message $m_i = \left(m_i^1, m_i^2, m_i^3, m_i^4\right)$, where $m_i^1 \in \Theta_i$, $m_i^2 \in \mathbb{Z}_+$, $m_i^3 : \Theta_{-i} \to Y$ with $m_i^3(\theta_{-i}) \in Y_i(\theta_{-i})$, $m_i^4 \in Y$. The outcome $g(m)$ is determined by the following rules:

Rule 1: If $m_i^2 = 1$ for all i, pick $f\left(m^1\right)$.

Rule 2: If there exists $j \in I$ such that $m_i^2 = 1$ for all $i \neq j$ and $m_j^2 > 1$, then pick $m_j^3\left(m_{-j}^1\right)$ with probability $1 - \frac{1}{m_j^2+1}$ and $\bar{y}_{m_{-j}^1}$ (as defined in 26) with probability $\frac{1}{m_j^2+1}$.

Rule 3: In all other cases, for each i, with probability $\frac{1}{I}\left(1 - \frac{1}{m_i^2+1}\right)$ pick m_i^4, and with probability $\frac{1}{I}\left(\frac{1}{m_i^2+1}\right)$ pick the interior lottery $\bar{y}$ (as defined in 25).

We first show that it is never a best reply for type θ_i to send a message with $m_i^2 > 1$ (i.e., $m_i \in b_i\left(\overline{S}\right) \Rightarrow m_i^2 = 1$). Suppose that θ_i has conjecture $\lambda_i \in \Delta\left(M_{-i} \times \Theta_{-i}\right)$. We can partition the messages of other agents as follows:

$$M_{-i}^*(\theta_{-i}) = \left\{m_{-i} : m_j^2 = 1 \text{ for all } j \neq i \text{ and } m_{-i}^1 = \theta_{-i}\right\},$$

and

$$\widehat{M}_{-i} = \left\{m_{-i} : m_j^2 > 1 \text{ for some } j \neq i\right\}.$$

By the conditional NTI property, we know that there exists $m_i^4 \in Y$ such that, if

$$\sum_{m_{-i}\in\widehat{M}_{-i},\theta_{-i}\in\Theta_{-i}} \lambda_i(m_{-i},\theta_{-i}) > 0,$$

then

$$\sum_{m_{-i}\in\widehat{M}_{-i},\theta_{-i}\in\Theta_{-i}} \lambda_i(m_{-i},\theta_{-i})\, u_i\left(m_i^4,\theta\right) > \sum_{m_{-i}\in\widehat{M}_{-i},\theta_{-i}\in\Theta_{-i}} \lambda_i(m_{-i},\theta_{-i})\, u_i(\bar{y},\theta).$$

And we also know from the conditional NTI property that there exists m_i^3 such that, if

$$\sum_{m_{-i}\in M_{-i}^*\left(\theta'_{-i}\right),\theta_{-i}\in\Theta_{-i}} \lambda_i(m_{-i},\theta_{-i}) > 0,$$

then

$$\sum_{m_{-i}\in M^*_{-i}\left(\theta'_{-i}\right),\theta_{-i}\in\Theta_{-i}} \lambda_i\left(m_{-i},\theta_{-i}\right) u_i\left(m_i^3\left(\theta'_{-i}\right),\theta\right) > \sum_{m_{-i}\in M^*_{-i}\left(\theta'_{-i}\right),\theta_{-i}\in\Theta_{-i}} \lambda_i\left(m_{-i},\theta_{-i}\right) u_i\left(\bar{y}_{\theta_{-i}},\theta\right).$$

Thus if $\left(m_i^1,m_i^2,m_i^3,m_i^4\right)$ with $m_i^2>1$ were a best response, then $\left(m_i^1,m_i^2+1,m_i^3,m_i^4\right)$ would be an even better response, contradiction.

Now fix any S with $m_i\in S_i\left(\theta_i\right)\Rightarrow m_i^2=1$. Let

$$\beta_i\left(\theta_i\right)=\left\{\theta'_i:\left(\theta'_i,1,m_i^3,m_i^4\right)\in S_i\left(\theta_i\right) \text{ for some } \left(m_i^3,m_i^4\right)\right\}.$$

First observe that EPIC implies that $\theta_i\in\beta_i\left(\theta_i\right)$. We will argue that if β is not acceptable, then $b\left(S\right)\neq S$. By robust monotonicity, we know that there exists i, θ_i, $\theta'_i\in\beta_i\left(\theta_i\right)$ such that, for all $\theta'_{-i}\in\Theta_{-i}$ and $\psi_i\in\Delta\left(\beta_{-i}^{-1}\left(\theta'_{-i}\right)\right)$, there exists $y\in Y_i\left(\theta'_{-i}\right)$ such that

$$\sum_{\theta_{-i}\in\Theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(y,\left(\theta_i,\theta_{-i}\right)\right) > \sum_{\theta_{-i}\in\Theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(f\left(\theta'_i,\theta'_{-i}\right),\left(\theta_i,\theta_{-i}\right)\right).$$

But now for any conjecture $\lambda_i\in\Delta\left(\left\{\left(m_{-i},\theta_{-i}\right):m_j^2=1 \text{ for all } j\neq i\right\}\right)$, there exists m_i^3 (with $m_i^3\left(\theta_{-i}\right)\in Y_i\left(\theta_{-i}\right)$) such that

$$\sum_{m_{-i},\theta_{-i}} \lambda_i\left(m_{-i},\theta_{-i}\right) u_i\left(m_i^3\left(m_{-i}^1\right),\theta\right) > \sum_{m_{-i},\theta_{-i}} \lambda_i\left(m_{-i},\theta_{-i}\right) u_i\left(f\left(\theta'_i,m_{-i}^1\right),\left(\theta_i,\theta_{-i}\right)\right).$$

Thus message $\left(\theta'_i,1,m_i^3,m_i^4\right)$ is never a best response for type θ_i.

We conclude that if

$$\beta_i\left(\theta_i\right)=\left\{\theta'_i:\left(\theta'_i,1,m_i^3,m_i^4\right)\in S_i^{\mathcal{M}}\left(\theta_i\right) \text{ for some } \left(m_i^3,m_i^4\right)\right\},$$

then β is acceptable. Thus f is materially rationalizably implemented.

Finally observe that $S^{\mathcal{M}}$ must satisfy the ex post best response property, with type θ_i sending a message of the form $\left(\theta'_i,1,m_i^3,m_i^4\right)$, so robust implementation is possible by Corollary 2. ■

We deliberately allowed for very badly behaved infinite mechanisms in order to make a tight connection with the existing literature and to get tight results. Many authors have argued that "integer game" constructions, like that we use in Theorem 1, should be viewed critically (see, e.g., Abreu and Matsushima (1992a) and Jackson (1992)). In our analysis of finite mechanisms in Section 3, the best responses were always well defined. As we saw there, the relationship between rationalizable and robust implementation is much simpler with the restriction to "nice" mechanisms, where best responses exists for all conjectures.

Part 1 of the above theorem represents a slight weakening of the necessary conditions of Propositions 2 and 3: semi-strict EPIC is weakened to EPIC and strict dual robust monotonicity is

weakened to dual robust monotonicity. These weaker conditions arise from allowing badly behaved mechanisms. Part 2 of the above theorem shows that they are also sufficient when combined with an no total indifference property.

The proof directly uses the link between rationalizable and robust implementation for the necessity as well as the sufficiency part. We briefly sketch the idea of the necessity part of the proof. If f is robustly implementable, then it is rationalizably implementable by Corollary 3. From rationalizable implementability, we then want to show that f satisfies strict robust monotonicity. We consider a given *and* unacceptable deception β. We start the process of iterative elimination and stop it at a specific round, denoted by k. This round k is the first round at which we can find an agent i, a true type profile θ_i and a report $\theta_i' \in \beta_i(\theta_i)$, such that a message, denoted by $\widehat{m}_i$, which will survive the process of iterated elimination for type θ_i', fails to survive the k-th round of elimination for type θ_i. We then show that the elimination of message $\widehat{m}_i$ at round k implies that the social choice function f satisfies strict robust monotonicity with respect to the deception β. Briefly, if $\widehat{m}_i$ survives the process of elimination for type θ_i', the message $\widehat{m}_i$ acts in the mechanism so as to report a payoff type θ_i'. If it is eliminated at round k for payoff type θ_i, then this means that for any belief agent i has over the remaining agents, there exists a message m_i^* which leads to an allocation through g which is strictly preferred by agent i when he has a payoff type θ_i. The significance of round k being the first round for which such an elimination relative to the deception β occurs, is that at this round, there do not yet exist any restrictions about message and payoff type profile regarding the other agents deception. The fact then that $\widehat{m}_i$ can be eliminated allows us to use full strength of the elimination argument to establish robust monotonicity. In the context of the proof it is interesting to note that the key step from iterative elimination to robust monotonicity is an argument which involves the early stages of the elimination process rather than the limit of iteration process.

6 Extensions, Variations and Discussion

6.1 Lotteries, Pure Strategies and Bayesian Implementation

In this section, we discuss how Theorem 1 is related to the classic literature on Bayesian implementation developed by Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989) and Jackson (1991). These authors asked whether it was possible to implement a social choice function in equilibrium on a fixed type space $\mathcal{T}$.[11] These authors analyzed the classic problem where attention was restricted to pure strategy equilibria and deterministic mechanisms. The assumption entails

[11] They allowed for more general social choice sets, but we restrict attention to functions for our comparison.

that the social choice function is a mapping $f : \Theta \to Z$ and the mechanism $g : M \to Z$. Note that in this classical approach it was not necessary to even define agent's preferences over lotteries and they certainly did not effect implementability.

Having fixed a type space, the natural notion of a pure strategy deception on the fixed type space is a collection $\alpha = (\alpha_1, ..., \alpha_I)$, with each $\alpha_i : T_i \to T_i$. Thus $\alpha : T \to T$ is defined by $\alpha(t) = (\alpha_i(t_i))_{i=1}^I$. The key monotonicity notion, translated into our language, is then the following:

Definition 15 (Bayesian Monotonicity)
Social choice function f satisfies Bayesian monotonicity on type space $\mathcal{T}$ if, for every deception α with $f\left(\widehat{\theta}(t)\right) \neq f\left(\widehat{\theta}(\alpha(t))\right)$ for some t, there exists i, t_i and $k : T \to Z$ such that

$$\sum_{t_{-i}} u_i\left(k(\alpha(t)), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_{-i})[t_i] > \sum_{t_{-i}} u_i\left(f\left(\widehat{\theta}(\alpha(t))\right), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_{-i})[t_i],$$

and

$$\sum_{t_{-i}} u_i\left(f\left(\widehat{\theta}(t_i', t_{-i})\right), \widehat{\theta}(t_i', t_{-i})\right) \widehat{\pi}_i(t_{-i})[t_i'] \geq \sum_{t_{-i}} u_i\left(k(\alpha_i(t_i), t_{-i}), \widehat{\theta}(t_i', t_{-i})\right) \widehat{\pi}_i(t_{-i})[t_i'], \ \forall t_i'.$$

Jackson shows that this condition is necessary for Bayesian implementation, and that a slight strengthening, Bayesian monotonicity no veto, is sufficient. We can also show that our robust monotonicity condition is equivalent to the requirement that Bayesian monotonicity is satisfied on all type spaces.

Proposition 8 (Equivalence)
Social choice function f satisfies Bayesian monotonicity on every type space if and only if it satisfies robust monotonicity.

The equivalence is established by a constructive proof via a specific type space. The constructive element is the identification of a type space on which Bayesian monotonicity is guaranteed to fail if robust monotonicity fails. It is worthwhile to note that the specific type space is much smaller than the universal type space. The proof of this result is in the appendix of the working paper version, Bergemann and Morris (2008b).

In some sense, the notion of robustness is more subtle in the context of full rather than partial implementation. With partial implementation, i.e. truthtelling in the direct mechanism, the universal type space is by definition the most difficult type space to obtain truthtelling. In the universal type space, every agent has the maximal number of possible misreports and hence the designer faces the maximal number of incentive constraints. In the context of full implementation,

the trade-off is ambiguous. As a larger type space contains by definition more types, it offers every agent more possibilities to misreport. But then, just as a larger type space made truthtelling more difficult to obtain, the other equilibria might also cease to exist after the introduction of additional types. This second part offers the possibility that larger type spaces facilitate rather than complicate the full implementation problem.

But note that this line of argument would establish the necessity of robust implementation if the planner is restricted to deterministic mechanisms (a disadvantage) but he can assume that agents follow pure strategies (an advantage). How do these assumptions matter?

First, observe that the advantage of restricting attention to pure strategies goes away completely when we require implementation on all type spaces: if there is a mixed strategy equilibrium that results in a socially sub-optimal outcome on some type space, we can immediately construct a larger type space (purifying the original equilibrium) where the socially sub-optimal outcome is played in a pure strategy equilibrium. Thus our robust analysis conveniently removes that unfortunate gap between pure and mixed strategy implementation that has plagued the implementation literature.

We use the extension to stochastic mechanisms in just two places. Ex post incentive compatibility and robust monotonicity would remain necessary conditions even if we restricted attention to deterministic mechanisms (the arguments would be unchanged). But, as we note in Footnote 10, even if lotteries were not used in the implementing mechanism, the implied robust monotonicity condition would involve lotteries (as rewards for whistle-blowers). But if lotteries were not allowed, our sufficiency argument would then require a slightly strengthened version of the robust monotonicity condition, with the lottery y replaced by a deterministic outcome. Our sufficiency argument also uses lotteries under Rules 1 and 2. As in a recent paper by Benoit and Ok (2008) on complete information implementation, we use lotteries to significantly weaken the sufficient conditions, so that we require only the conditional NTI property in addition to EPIC and robust monotonicity. If we did not allow lotteries in this part of the argument, we would require a much stronger economic condition in the spirit of Jackson's "Bayesian monotonicity no veto" condition. We have developed combined robust monotonicity and economic conditions (not reported here) sufficient for interim implementation on all full support types spaces. However, an additional complication is that, without lotteries in the implementing mechanism, we cannot establish sufficiency on type spaces where agents have disjoint supports.

It is possible to construct a simple example where EPIC and robust monotonicity are not sufficient for robust monotonicity without lotteries by taking the coordination example of Section 3.6 but removing the outcomes z and z'. As we show in the Appendix (Section 7.3), robust implementation is then not possible in this example despite the fact that the social choice function selects a unique strictly Pareto-dominant outcome at every type profile.

6.2 Ex Post and Robust Implementation

In contrast to our earlier results in Bergemann and Morris (2005b), where we showed that robust *partial* implementation is equivalent to ex post incentive compatibility, robust implementation is in general a more demanding notion of implementation than ex post equilibrium implementation. The following simple example, introduced by Palfrey and Srivastava (1989), is useful to relate the different implementation notions and understand the role of interdependent types. In this example, there are three agents and each agent has two possible "payoff types", θ_a or θ_b. There are two possible choices for society, a or b. All agents have identical preferences. If a majority of agents (i.e., at least two) are of type θ_y, then every agent gets utility 1 from outcome y and utility 0 from the other outcome. The social choice function agrees with the common preferences of the agents. Thus $f : \{\theta_a, \theta_b\}^3 \to \{a, b\}$ satisfies $f(\theta) = y$ if and only if $\#\{i : \theta_i = \theta_y\} \geq 2$.

Clearly, ex post incentive compatibility is not a problem in this example. The problem is that in the "direct mechanism" - where all agents simply announce their types - there is the possibility that all agents will choose to always announce θ_a. Since no agent expects to be pivotal, he has no incentive to truthfully announce his type when he is in fact θ_b. What happens if we allow more complicated mechanisms?

If there were complete information about agents' preferences, then the social choice function is clearly implementable: the social planner could pick an agent, say agent 1, and simply follow that agent's recommendation.

But suppose instead that there is incomplete information about agents' preferences. In particular, suppose it is common knowledge that each agent's type is θ_b with independent probability q, with $q^2 > \frac{1}{2}$. This example fails the Bayesian monotonicity condition of Postlewaite and Schmeidler (1986) and Jackson (1991). Palfrey and Srivastava (1989) observe that it is also not possible to implement in undominated Bayesian Nash equilibrium in this example.

Bergemann and Morris (2008a) have analyzed the alternative "more robust" solution concept of ex post equilibrium in this context. It is easy to construct an augmented mechanism whose only ex post equilibrium delivers the social choice function. Let each agent send a message $m_i \in \{\theta_a, \theta_b\} \times \{\text{truth, lie}\}$, with the interpretation that an agent is announcing his own type and also sends the message "truth" if he thinks that others are telling the truth and sends the message "lie" if he thinks that someone is lying. Outcome y is implemented if a majority claim to be type θ_y and all agents announce "truth"; or if either 1 or 3 agents claim to be type θ_y and at least one agent reports lying.

There is a truthtelling ex post equilibrium where each agent truthfully announces his type and also announces "truth". Now suppose there exists an ex post equilibrium such that at some type

profile, the desired outcome is not chosen. Note that whatever the announcements of the other agents, each agent always has the ability to determine the outcome y, by sending the message "lie" and - given the announcements of the other agents - choosing his message so that an odd number of agents have claimed to be type θ_y. So this is not consistent with ex post equilibrium.

Robust implementation is impossible in this example. Consider the type space where there is common knowledge that whenever an agent is type θ_y, he assigns probability $\frac{1}{2}$ to both of the other agents being type $y' \neq y$ and probability $\frac{1}{2}$ to one being type y and the other being y'. Thus every type of every agent thinks there is a 50% chance that outcome a is better and a 50% chance that b is better. Evidently, there is no way of designing a mechanism that ensures that agents do not fully pool. But if they fully pool, robust implementation is not possible.

6.3 Extensions

The previous sections examined the importance of our assumptions about lotteries over outcomes and restrictions on mechanisms. We also restricted attention in our main analysis to the case of discrete but infinite pure outcomes Z, payoff types Θ_i and types T_i. While most of our results would extend naturally to more general Z, Θ_i and T_i, the formal treatment of non-compact type spaces would raise technical issues that we have chosen to avoid.

7 Appendix A

7.1 Robust Monotonicity and Dual Robust Monotonicity

Lemma 2

1. *If f satisfies (strict) robust monotonicity, then f satisfies dual (strict) robust monotonicity.*
2. *If deterministic outcomes and payoff types are finite, and f satisfies dual (strict) robust monotonicity, then f satisfies (strict) robust monotonicity.*

Proof. (1.) follows immediately from the definitions. To prove (2.), suppose that outcomes and payoff types are finite and that f satisfies dual (strict) robust monotonicity. Then for every unacceptable deception β, there exist i, θ_i, $\theta_i' \in \beta_i(\theta_i)$ such that, for all $\theta_{-i}' \in \Theta_{-i}$, there exists a compact set $\overline{Y} \subseteq Y$ such that $y \in \overline{Y}$ implies

$$u_i\left(f\left(\theta_i'', \theta_{-i}\right), \left(\theta_i', \theta_{-i}'\right)\right) \geq (>)\, u_i\left(y, \left(\theta_i'', \theta_{-i}'\right)\right)$$

for all θ_i'' with $f\left(\theta_i'', \theta_{-i}\right) \neq y$ and, for each $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right)$, there exists $y \in \overline{Y}$ such that

$$\sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i(\theta_{-i})\, u_i(y, (\theta_i, \theta_{-i})) > \sum_{\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)} \psi_i(\theta_{-i})\, u_i\left(f\left(\theta_i', \theta_{-i}'\right), (\theta_i, \theta_{-i})\right).$$

By the equivalence between strict domination and never a best response (see Theorem 2.10 in Gale (1989)), we have that there exists $y^* \in \overline{Y}$ with

$$u_i(y^*, (\theta_i, \theta_{-i})) > u_i\left(f\left(\theta_i', \theta_{-i}'\right), (\theta_i, \theta_{-i})\right)$$

for all $\theta_{-i} \in \beta_{-i}^{-1}\left(\theta_{-i}'\right)$. This establishes (strict) robust monotonicity. ■

7.2 A Badly Behaved Mechanism

The example illustrates the gap between the necessary and sufficient conditions in Proposition 7. Specifically, it shows that there can be an equilibrium for every type space $\mathcal{T}$ in a mechanism, yet $S^{\mathcal{M}}$ does not satisfy the ex post best response property.

In the example, there are two agents and there is complete information, so each agent has a unique type. There are a finite number of outcomes $Z = \{a, b, c\}$. The payoffs are given by the following table:

	a	b	c
agent 1	0	-1	$+1$
agent 2	0	0	0

The planner's choice (in the unique payoff state) is a. Thus it is trivial to robustly implement the social choice function. But suppose that the planner chooses the following (strange) mechanism: $M_1 = \{1, 2, 3, ...\}$, $M_2 = \{1, 2\}$ and

$$g(m_1, m_2) = \begin{cases} a, & \text{if} \quad m_1 = 1 \\ b, & \text{if} \quad m_1 > 1 \text{ and } m_2 = 1 \\ \left[\frac{1}{m_1}, b; \left(1 - \frac{1}{m_1}\right), c\right], & \text{if} \quad m_1 > 1 \text{ and } m_2 = 2 \end{cases}$$

where $\left[\frac{1}{m_1}, b; \left(1 - \frac{1}{m_1}\right), c\right]$ is the lottery putting probability $\frac{1}{m_1}$ on b and probability $\left(1 - \frac{1}{m_1}\right)$ on c. Thus $g(m_1, m_2)$ can be represented by the following table:

g	1	2
1	a	a
2	b	$\left[\frac{1}{2}, b; \frac{1}{2}, c\right]$
3	b	$\left[\frac{1}{3}, b; \frac{2}{3}, c\right]$
$\vdots$	$\vdots$	$\vdots$
k	b	$\left[\frac{1}{k}, b; 1 - \frac{1}{k}, c\right]$
$\vdots$	$\vdots$	$\vdots$

Thus the agents are playing the following complete information game:

m_1/m_2	1	2
1	$0, 0$	$0, 0$
2	$-1, 0$	$0, 0$
3	$-1, 0$	$\frac{1}{3}, 0$
$\vdots$	$\vdots$	$\vdots$
k	$-1, 0$	$1 - \frac{2}{k}, 0$
$\vdots$	$\vdots$	$\vdots$

Now on any type space, there is always an equilibrium where agent 1 chooses action 1 and agent 2 chooses action 1, and outcome a is chosen. Moreover, on any type space, in any equilibrium, outcome a is always chosen: if agent 1 ever has a best response not to play 1 then he has no best response. So he always plays 1 in equilibrium. Thus the trivial social choice function is robustly implemented by this mechanism.

While only message 1 survives iterated deletion of never best responses for agent 1, both messages survive iterated deletion of never best responses for agent 2. Thus we have $S_1^{\mathcal{M}} = \{1\}$ and $S_2^{\mathcal{M}} = \{1, 2\}$. Note that $S^{\mathcal{M}}$ satisfies the interim best response property, see Definition 13, but not

the ex post best response property, see Definition 12. For we observe that

$$u_1(g(1,2)) = u_1(a) = 0 < \frac{1}{2} = u_1(g(2,2)),$$

violating the ex post best response property.

The insight of the example is that the quantifier "for every type space $\mathcal{T}$" does not necessarily guarantee that all actions which will be chosen with positive probability in some equilibrium and for some type space, will also be chosen with probability one in some equilibrium for some type space. For this reason, the quantifier "for every type space $\mathcal{T}$" does not allow us to establish a local, i.e. ex post best response property of every action in $S^{\mathcal{M}}$.

7.3 Coordination Example Continued

The final example is the pure coordination game, which we first considered in Section 3.6, but without the additional allocations, z and z'. It illustrates the importance of lotteries for robust implementation. The example will satisfy EPIC and robust monotonicity, yet it cannot be robustly implemented without the use of lotteries. On the other hand the preferences clearly satisfy the conditional NTI property, and hence the sufficient conditions for robust implementation would be satisfied with lotteries.

As in the example in Section 3.6, the payoffs of the player are given by (15) and the social choice function f is given by (16). The social choice function is strictly ex post incentive compatible but there is another equilibrium in the "direct mechanism" where each agent misreports his type, and each agent gets a payoff of 1.

Robust monotonicity is clearly satisfied even if the rewards $Y_i(\theta_{-i})$ are restricted to the deterministic allocations Z. We will show that robust implementation is not possible even in an infinite mechanism if we restrict attention to deterministic mechanisms. Fix a mechanism $\mathcal{M}$. Let

$$S_i^*(\theta_i) = \{m_i : g(m_i, m_j) = f(\theta_i, \theta_j) \text{ for some } m_j, \theta_j\},$$

be the set of messages for agent i which would select the allocation recommended by the social choice function for some m_j, θ_j. We now show by induction that, $S_i^*(\theta_i) \subseteq S_i^k(\theta_i)$ for all k using the structure of the payoffs. Suppose that this is true for k. Then for any $m_i \in S_i^*(\theta_i) \subseteq S_i^k(\theta_i)$, there exists $m_j \in S_j^*(\theta_j) \subseteq S_j^k(\theta_j)$ such that $g(m_i, m_j) = f(\theta_i, \theta_j)$. Thus there does not exist $\nu_i \in \Delta(M_i)$ such that

$$\sum_{m_i'} \nu_i(m_i') u_i(g(m_i', m_j), (\theta_i, \theta_j)) > u_i(g(m_i, m_j), (\theta_i, \theta_j)) = 3.$$

So $m_i \in S_i^{k+1}(\theta_i)$.

Thus we must have that $(m_1, m_2) \in S_1^*(\theta_1) \times S_2^*(\theta_2)$ implies $g(m_1, m_2) = f(\theta_1, \theta_2)$. Let $m_i^*(\cdot)$ be any selection from $S_i^*(\cdot)$. Now let k^* be the lowest k such that, for some i,

$$m_i^*\left(\theta_i'\right) \notin S_i^k(\theta_i).$$

Without loss of generality, let $i = 1$. Note $m_2^*\left(\theta_2'\right) \in S_2^{k-1}(\theta_2)$ by definition of k^*. If agent 1 was type θ_1 and was sure his opponent were type θ_2 and choosing action $m_2^*\left(\theta_2'\right)$, we know that he could guarantee himself a payoff of 1 by choosing $m_1^*\left(\theta_1'\right)$. Since $m_1^*\left(\theta_1'\right)$ is deleted for type θ_1 at round k^*, we know that there exists $\nu_1 \in \Delta(M_1)$ such that

$$\sum_{m_1'} \nu_1\left(m_1'\right) g_1\left(m_1', m_2^*\left(\theta_2'\right)\right) > 1,$$

and thus there exists m_1' such that $g_1\left(m_1', m_2^*\left(\theta_2'\right)\right) = f(\theta_1, \theta_2)$. This implies that $m_2^*\left(\theta_2'\right) \in S_2^*(\theta_2)$, a contradiction.

The example uses the fact that the social choice function always selects an outcome that is strictly Pareto-optimal and - paradoxically - it is this feature which inhibits rationalizable implementation in the current example. Borgers (1995) proves the impossibility of complete information implementation of non-dictatorial social choice functions in iteratively undominated strategies when the set of feasible preference profiles includes such unanimous preference profiles and the argument here is reminiscent of Borgers' argument.

8 Appendix B

8.1 Robust Monotonicity and Robust Measurability

In this section we document that robust measurability neither implies nor is implied by robust monotonicity. This observation parallels an observation in the standard Bayesian implementation literature. Abreu and Matsushima (1992b) showed that Bayesian incentive compatibility and a measurability conditions, henceforth referred to as Abreu-Matsushima measurability are necessary conditions for virtual implementation in Bayesian Nash equilibrium in well-behaved mechanisms. Serrano and Vohra (2005) describe a "virtual monotonicity" condition - a weakening of the Bayesian monotonicity condition of Postlewaite and Schmeidler (1986) and Jackson (1991) - which, together with Bayesian incentive compatibility, is necessary and sufficient for virtual implementation in Bayesian Nash equilibrium using perhaps badly behaved mechanisms. Virtual monotonicity must therefore be a weakening of Abreu-Matsushima measurability. Example 2 in Serrano and Vohra (2001) exhibits an environment where all non-constant social choice functions fail Abreu-Matsushima measurability fails but all social choice functions satisfy virtual monotonicity and many satisfy Bayesian monotonicity. On the other hand, the social choice function allocating a single object efficiently under private values will fail Bayesian monotonicity (any efficient allocation mechanism will allow undesirable equilibria) but will satisfy Abreu-Matsushima measurability. Thus Bayesian monotonicity neither implies nor is implied by Abreu-Matsushima measurability.

Example 1: Robust Measurability holds while Robust Monotonicity fails Consider an environment with two agents, a and b. The payoff type space of each agent i is given by $\Theta_i = \left\{\theta_i^1, \theta_i^2, \theta_i^3\right\}$. The allocation space is given by $X = \{x_1, x_2, x_3, x_4\}$. We display below the payoff of agent a. The payoffs for agent b are symmetric and are obtained by switching the subscripts a and b below:

x_1	θ_b^1	θ_b^2	θ_b^3
θ_a^1	1	1	1
θ_a^2	1	1	0
θ_a^3	0	0	1

x_2	θ_b^1	θ_b^2	θ_b^3
θ_a^1	0	0	1
θ_a^2	1	1	1
θ_a^3	1	1	0

x_3	θ_b^1	θ_b^2	θ_b^3
θ_a^1	1	1	0
θ_a^2	0	0	1
θ_a^3	1	1	1

x_4	θ_b^1	θ_b^2	θ_b^3
θ_a^1	$\frac{2}{3}+\varepsilon$	$\frac{2}{3}-\varepsilon$	$\frac{2}{3}$
θ_a^2	$\frac{2}{3}+\varepsilon$	$\frac{2}{3}-\varepsilon$	$\frac{2}{3}$
θ_a^3	$\frac{2}{3}+\varepsilon$	$\frac{2}{3}-\varepsilon$	$\frac{2}{3}$

. (27)

Given the payoff matrix given by (27), it is immediate to show that the limit set of separable set, Ξ_i^*, as defined earlier in (11) is given by the collection of singleton sets with $\Xi_i^* = \left\{\left\{\theta_i^1\right\}, \left\{\theta_i^2\right\}, \left\{\theta_i^3\right\}\right\}$. The limit set is obtained by observing that a subset Ψ_j separates a subset Ψ_i whenever $\#\Psi_i \geq \#\Psi_j$ and $\#\Psi_i > 1$ but not if $\#\Psi_i < \#\Psi_j$ or $\#\Psi_i = 1$. Thus while at the 0-th

level, the inseparable sets of agent i, Ξ_i^0, will consist of all subsets of payoff types, the 1-st level of inseparable sets of agent i, Ξ_i^1, will only consist of all sets of payoff types with cardinality at most 2, and the 2-nd level of inseparable sets of agent i, Ξ_i^2, will consist exactly of all singletons. It follows that the robust measurability condition does not impose any restrictions on the social choice functions.

We illustrate the process of separation for $\Psi_a = \Theta_a$ and $\Psi_b = \Theta_b$. A given type θ_a has a belief $\lambda_{\theta_a}(\theta_b)$ over the payoff types of agent b. We denote $p_{\theta_a} \triangleq \lambda_{\theta_a}\left(\theta_b^1\right)$ and $q_{\theta_a} \triangleq \lambda_{\theta_a}\left(\theta_b^2\right)$. The expected utility of type θ_a with belief $\lambda_{\theta_a}(\theta_b)$ over the allocations $\{x_1, x_2, x_3, x_4\}$ is then given by for the three types $\theta_a^1, \theta_a^2, \theta_a^3$ by:

$$\begin{array}{lccccc} (& 1, & 1 - p_{\theta_a^1} - q_{\theta_a^1}, & p_{\theta_a^1} + q_{\theta_a^1}, & \frac{2}{3} + \varepsilon\left(p_{\theta_a^1} - q_{\theta_a^1}\right) &) \\ (& p_{\theta_a^2} + q_{\theta_a^2}, & 1, & 1 - p_{\theta_a^2} - q_{\theta_a^2}, & \frac{2}{3} + \varepsilon\left(p_{\theta_a^2} - q_{\theta_a^2}\right) &) \\ (& 1 - p_{\theta_a^3} - q_{\theta_a^3}, & p_{\theta_a^3} + q_{\theta_a^3}, & 1, & \frac{2}{3} + \varepsilon\left(p_{\theta_a^3} - q_{\theta_a^3}\right) &) \end{array} \tag{28}$$

It is now immediate to verify that there does not exist a triple of beliefs $\left(p_{\theta_a^1}, q_{\theta_a^1}\right)$, $\left(p_{\theta_a^2}, q_{\theta_a^2}\right)$ and $\left(p_{\theta_a^3}, q_{\theta_a^3}\right)$ such that the cardinal preference profiles of the payoff types θ_a^1, θ_a^2 and θ_a^3 coincide. In fact, for every pair among the three types, we can find identical preference profiles, but not for the triple itself.

We next show that the robust monotonicity condition fails for some social choice functions f. In fact, consider the following ex post incentive compatible social choice function f given by:

f	θ_b^1	θ_b^2	θ_b^3
θ_a^1	x_1	x_2	x_2
θ_a^2	x_2	x_2	x_2
θ_a^3	x_2	x_2	x_3

(29)

For the given payoff environment (27) and the social choice function (29), we consider the deception with $\beta_i(\theta_i) = \Theta_i$ for all i and θ_i. By symmetry of the payoffs and the deceptions across states, it suffices to consider a single type profile, namely $\theta_i = \theta_a^1$ and $\theta_i' = \theta_a^2$ and likewise $\theta_{-i}' = \theta_b^2$. The robust monotonicity conditions require that there exists a reward y such that:

$$\begin{array}{rcl} u_a\left(y, \left(\theta_a^1, \theta_b^1\right)\right) & > & u_a\left(f\left(\theta_a^2, \theta_b^2\right), \left(\theta_a^1, \theta_b^1\right)\right), \\ u_a\left(y, \left(\theta_a^1, \theta_b^2\right)\right) & > & u_a\left(f\left(\theta_a^2, \theta_b^2\right), \left(\theta_a^1, \theta_b^2\right)\right), \\ u_a\left(y, \left(\theta_a^1, \theta_b^3\right)\right) & > & u_a\left(f\left(\theta_a^2, \theta_b^2\right), \left(\theta_a^1, \theta_b^3\right)\right), \end{array} \tag{30}$$

and

$$\begin{aligned} u_a\left(f\left(\theta_a^1,\theta_b^2\right),\left(\theta_a^1,\theta_b^2\right)\right) &\geq u_a\left(y,\left(\theta_a^1,\theta_b^2\right)\right), \\ u_a\left(f\left(\theta_a^2,\theta_b^2\right),\left(\theta_a^2,\theta_b^2\right)\right) &\geq u_a\left(y,\left(\theta_a^2,\theta_b^2\right)\right), \\ u_a\left(f\left(\theta_a^3,\theta_b^2\right),\left(\theta_a^3,\theta_b^2\right)\right) &\geq u_a\left(y,\left(\theta_a^3,\theta_b^2\right)\right). \end{aligned} \tag{31}$$

As we evaluate the inequalities (30) for the given payoff environment (27) and the social choice function (29), we find that the values on the right hand side are given as follows:

$$\begin{aligned} u_a\left(y,\left(\theta_a^1,\theta_b^1\right)\right) &> u_a\left(x_2,\left(\theta_a^1,\theta_b^1\right)\right)=0, \\ u_a\left(y,\left(\theta_a^1,\theta_b^2\right)\right) &> u_a\left(x_2,\left(\theta_a^1,\theta_b^2\right)\right)=0, \\ u_a\left(y,\left(\theta_a^1,\theta_b^3\right)\right) &> u_a\left(x_2,\left(\theta_a^1,\theta_b^3\right)\right)=1. \end{aligned}$$

But as the value of the last value of the misreport is 1 for agent a, it follows that we cannot find a reward which strictly exceeds 1. It follows that the robust monotonicity condition is violated in this example.

The difference between robust measurability and robust monotonicity here stems from the fact the different types of agent i could be separated in the payoff environment with allocations which were not called upon by the social choice function. The robust monotonicity condition was therefore limited to use the allocation x_4 as a reward, but here it failed to provide a higher utility than the one provided by the social choice function under some report profiles.

Example 2: Robust Measurability fails but Robust Monotonicity holds There are two agents, a and b, and each agent i has two types, θ_i and θ_i'. There are six pure outcomes, $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$. The state dependent utility of agents a and b are depicted in the following tables:

x_1	θ_b	θ_b'
θ_a	$2,2$	$-2,0$
θ_a'	$0,0$	$0,2$

x_2	θ_b	θ_b'
θ_a	$-2,0$	$2,2$
θ_a'	$0,2$	$0,0$

x_3	θ_b	θ_b'
θ_a	$0,0$	$0,2$
θ_a'	$2,2$	$-2,0$

x_4	θ_b	θ_b'
θ_a	$0,2$	$0,0$
θ_a'	$-2,0$	$2,2$

(32)

and

x_5	θ_b	θ_b'
θ_a	$0,0$	$0,0$
θ_a'	$0,0$	$0,0$

x_6	θ_b	θ_b'
θ_a	$1,1$	$1,1$
θ_a'	$1,1$	$1,1$

(33)

Suppose agent a assigns equal probability to each type of agent b. Then - whatever his payoff type - his expected utility from allocations $(x_1, x_2, x_3, x_4, x_5, x_6)$ are $(0, 0, 0, 0, 0, 1)$. Thus the set

$\Theta_b = \left\{\theta_b, \theta_b'\right\}$ does not separate the set $\Theta_a = \left\{\theta_a, \theta_a'\right\}$. A symmetric argument establishes that Θ_a does not separate Θ_b. We conclude that every pair of types of each agent is not separable and hence only constant social choice functions satisfy robust measurability.

But consider the following (Pareto-efficient) social choice function:

f	θ_b	θ_b'
θ_a	x_1	x_2
θ_a'	x_3	x_4

(34)

Given the social choice function f, it is immediate to verify that strict ex post incentive compatibility and robust monotonicity both hold. To verify the latter, observe that consider a deception with $\beta_i\left(\widetilde{\theta}_i\right) = \Theta_i$ for some $\widetilde{\theta}_i$. Without loss of generality, assume that $\beta_a\left(\theta_a\right) = \Theta_a$. Now type θ_a reporting θ_a' can be offered outcome x_6 to report the deception. Given the definition of robust monotonicity, we need to verify that

$$\begin{array}{c} u_a\left(y, \left(\theta_a, \theta_b\right)\right) > u_a\left(f\left(\theta_a', \theta_b'\right), \left(\theta_a, \theta_b\right)\right) \\ u_a\left(y, \left(\theta_a, \theta_b'\right)\right) > u_a\left(f\left(\theta_a', \theta_b'\right), \left(\theta_a, \theta_b'\right)\right) \end{array} \tag{35}$$

as well as

$$\begin{array}{c} u_a\left(f\left(\theta_a, \theta_b'\right), \left(\theta_a, \theta_b'\right)\right) \geq u_a\left(y, \left(\theta_a, \theta_b'\right)\right) \\ u_a\left(f\left(\theta_a', \theta_b'\right), \left(\theta_a', \theta_b'\right)\right) \geq u_a\left(y, \left(\theta_a', \theta_b'\right)\right) \end{array}. \tag{36}$$

Given the payoff described in (32) and (33) and given the social choice function f, we can verify that with $y = x_6$, we have:

$$\begin{array}{c} 1 = u_a\left(y, \left(\theta_a, \theta_b\right)\right) > u_a\left(f\left(\theta_a', \theta_b'\right), \left(\theta_a, \theta_b\right)\right) = 0 \\ 1 = u_a\left(y, \left(\theta_a, \theta_b'\right)\right) > u_a\left(f\left(\theta_a', \theta_b'\right), \left(\theta_a, \theta_b'\right)\right) = 0 \end{array}$$

and

$$\begin{array}{c} 2 = u_a\left(f\left(\theta_a, \theta_b'\right), \left(\theta_a, \theta_b'\right)\right) \geq u_a\left(y, \left(\theta_a, \theta_b'\right)\right) = 1 \\ 2 = u_a\left(f\left(\theta_a', \theta_b'\right), \left(\theta_a', \theta_b'\right)\right) \geq u_a\left(y, \left(\theta_a', \theta_b'\right)\right) = 1 \end{array}.$$

The social choice function f can be robustly implemented with the following mechanism. Agent a sends a message $m_a \in M_a = \left\{\theta_a, \theta_a'\right\} \cup \{1, 2, 3, ...\}$, agent b sends a message $m_b \in M_b = \left\{\theta_b, \theta_b'\right\}$. If $m_a \in \left\{\theta_a, \theta_a'\right\}$, then $g\left(m_a, m_b\right) = f\left(m_a, m_b\right)$; if $m_a \in \{1, 2, 3, ...\}$, then $g\left(m_a, m_b\right)$ is the lottery putting probability $\frac{1}{m_a}$ on x_5 and probability $1 - \frac{1}{m_a}$ on x_6. Now truth-telling survives iterated deletion of never best responses. Also, (i) sending message 2 with an expected payoff of $\frac{1}{2}$ is always a better response for agent a than misreporting his type with a payoff 0, given thruthtelling by agent b and (ii) choosing $m_a + 1$ with an expected payoff of $\frac{m_a}{m_a+1}$ is always a better response for agent a than sending message m_a with an expected payoff: $\frac{m_a-1}{m_a}$. So player a must tell the truth and truth-telling is then the only best response for agent b.

8.2 Bayesian Monotonicity

This subsection contains the proof of Proposition 8 which establishes the equivalence between robust monotonicity and Bayesian monotonicity on every type space by means of a constructive proof (via a specific type space).

Proof of Proposition 8. ($\Rightarrow$) We will show that if robust monotonicity fails, we can construct a type space where Bayesian monotonicity fails. The argument will be constructive.

Fix an unacceptable deception β. Suppose that robust monotonicity fails. Then for each i, θ_i, $\theta_i' \in \beta_i(\theta_i)$, there exist

$$\theta_{-i}\left[\theta_i, \theta_i'\right] \in \Theta_{-i} \quad \text{and} \quad \psi_i\left[\theta_i, \theta_i'\right] \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}\left[\theta_i, \theta_i'\right]\right)\right) \tag{37}$$

such that:

$$u_i\left(f\left(\theta_i'', \theta_{-i}\left[\theta_i, \theta_i'\right]\right), \left(\theta_i'', \theta_{-i}\left[\theta_i, \theta_i'\right]\right)\right) \geq u_i\left(y, \left(\theta_i'', \theta_{-i}\left[\theta_i, \theta_i'\right]\right)\right), \quad \forall \theta_i'' \in \Theta_i \tag{38}$$

implies

$$\sum_{\theta_{-i} \in \Theta_{-i}} \psi_i(\theta_{-i})\left[\theta_i, \theta_i'\right] u_i\left(f\left(\theta_i', \theta_{-i}\left[\theta_i, \theta_i'\right]\right), (\theta_i, \theta_{-i})\right) \geq \sum_{\theta_{-i} \in \Theta_{-i}} \psi_i(\theta_{-i})\left[\theta_i, \theta_i'\right] u_i(y, (\theta_i, \theta_{-i})). \tag{39}$$

Now we construct a type space around θ_i, θ_i' and $\psi_i\left[\theta_i, \theta_i'\right]$ given by (37) for which Bayesian monotonicity fails. First, agent i has a set of "deception" types T_i^1 which are isomorphic to $\Xi_i = \left\{\left(\theta_i, \theta_i'\right) : \theta_i \in \Theta_i \text{ and } \theta_i' \in \beta_i(\theta_i)\right\}$; thus there exists a bijection $\xi_i^1 : T_i^1 \to \Xi_i$. The type responding to $\left(\theta_i, \theta_i'\right)$ has payoff type θ_i and believes that the other agents are of type:

$$\left(\left[\xi_j^1\right]^{-1}\left(\theta_j, \theta_{ij}\left[\theta_i, \theta_i'\right]\right)\right)_{j \neq i}$$

with probability $\psi_i(\theta_{-i})\left[\theta_i, \theta_i'\right]$. Second, agent i has a set of "pseudo-complete information types" T_i^2, which are isomorphic to Θ; thus there exists a bijection $\xi_i^2 : T_i^2 \to \Theta$. The type corresponding to θ has payoff type θ_i and he is convinced that each other agent j is type $\left[\xi_j^1\right]^{-1}(\theta_j, \theta_j)$.

Slightly more formally, we have

$$T_i = T_i^1 \cup T_i^2.$$

If $t_i \in T_i^1$ and $\xi_i^1(t_i) = \left(\theta_i, \theta_i'\right)$, then

$$\widehat{\theta}_i(t_i) = \theta_i;$$

if $t_i \in T_i^2$ and $\xi_i^2(t_i) = \theta$, then

$$\widehat{\theta}_i(t_i) = \theta_i.$$

If $t_i \in T_i^1$ and $\xi_i^1(t_i) = \left(\theta_i, \theta_i'\right)$, then

$$\pi_i^*(t_{-i})[t_i] = \begin{cases} \psi_i(\theta_{-i})\left[\theta_i, \theta_i'\right], & \text{if} \quad t_{-i} \in T_{-i}^1 \text{ and } \theta_{-i} = \left(\left[\xi_j^1\right]^{-1}\left(\theta_j, \theta_{ij}\left[\theta_i, \theta_i'\right]\right)\right)_{j \neq i} \\ 0, & \text{if} \qquad \text{otherwise} \end{cases}$$

If $t_i \in T_i^2$ and $\xi_i^2(t_i) = \theta$, then

$$\pi_i^*(t_{-i})[t_i] = \begin{cases} 1, & \text{if} \quad t_{-i} \in T_{-i}^1 \text{ and } \theta_{-i} = \left(\left[\xi_j^1\right]^{-1}\left(\theta_j, \theta_{ij}\left[\theta_i, \theta_i'\right]\right)\right)_{j \neq i} \\ 0, & \text{if} \qquad \text{otherwise} \end{cases}$$

Now consider the Bayesian deception on this type space where each type $\left[\xi_i^1\right]^{-1}\left(\theta_i, \theta_i'\right)$ reports himself to be type $\left[\xi_i^1\right]^{-1}\left(\theta_i', \theta_i'\right)$, and all other types report their types truthfully. Thus

$$\alpha_i(t_i) = \begin{cases} \left[\xi_i^1\right]^{-1}\left(\theta_i', \theta_i'\right), & \text{if} \quad t_i = \left[\xi_i^1\right]^{-1}\left(\theta_i, \theta_i'\right) \\ t_i, & \text{if} \qquad \text{otherwise} \end{cases}.$$

Since β was unacceptable, we must have that $f\left(\widehat{\theta}(t)\right) \neq f\left(\widehat{\theta}(\alpha(t))\right)$ for some t. Thus the Bayesian monotonicity condition (Definition 15) for this type space requires that there exist i, t_i and $h : T \to Z$ such that

$$\sum_{t_{-i} \in T_{-i}} u_i\left(h(\alpha(t)), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_{-i})[t_i] > \sum_{t_{-i} \in T_{-i}} u_i\left(f\left(\widehat{\theta}(\alpha(t))\right), \widehat{\theta}(t)\right) \widehat{\pi}_i(t_{-i})[t_i], \tag{40}$$

and

$$\begin{aligned} & \sum_{t_{-i} \in T_{-i}} u_i\left(f\left(\widehat{\theta}\left(t_i'', t_{-i}\right)\right), \widehat{\theta}\left(t_i'', t_{-i}\right)\right) \widehat{\pi}_i(t_{-i})\left[t_i''\right] \\ \geq & \sum_{t_{-i} \in T_{-i}} u_i\left(h\left(\alpha_i(t_i), t_{-i}\right), \widehat{\theta}\left(t_i'', t_{-i}\right)\right) \widehat{\pi}_i(t_{-i})\left[t_i''\right], \ \forall t_i''. \end{aligned} \tag{41}$$

The t_i cannot be an element of T_i^2, because such a type does not expect any deviation from truth-telling under the deception. So it must be an element of T_i^1, with $\xi_i^1(t_i) = \left(\theta_i, \theta_i'\right)$. Now condition (40) becomes

$$\begin{aligned} & \sum_{\theta_{-i} \in \Theta_{-i}} u_i\left(h\left(\left[\xi_i^1\right]^{-1}\left(\theta_i', \theta_i'\right), \left(\left(\left[\xi_j^1\right]^{-1}\left(\theta_{ij}\left[\theta_i, \theta_i'\right], \theta_{ij}\left[\theta_i, \theta_i'\right]\right)\right)_{j \neq i}\right)\right), (\theta_i, \theta_{-i})\right) \psi_i(\theta_{-i})\left[\theta_i, \theta_i'\right] \\ > & \sum_{\theta_{-i} \in \Theta_{-i}} u_i\left(f\left(\theta_i', \theta_{-i}\left[\theta_i, \theta_i'\right]\right), (\theta_i, \theta_{-i})\right) \psi_i(\theta_{-i})\left[\theta_i, \theta_i'\right]. \end{aligned} \tag{42}$$

But letting t_i'' in condition (41) be in T_i^2 with $\xi_i^2\left(t_i''\right) = \left(\theta_i'', \theta_{-i}\left[\theta_i, \theta_i'\right]\right)$, we have

$$\begin{aligned} & u_i\left(f\left(\theta_i'', \theta_{-i}\left[\theta_i, \theta_i'\right]\right), \left(\theta_i'', \theta_{-i}\left[\theta_i, \theta_i'\right]\right)\right) \\ \geq & u_i\left(h\left(\left[\xi_i^1\right]^{-1}\left(\theta_i', \theta_i'\right), \left(\left[\xi_j^1\right]^{-1}\left(\theta_{ij}\left[\theta_i, \theta_i'\right], \theta_{ij}\left[\theta_i, \theta_i'\right]\right)\right)_{j \neq i}\right), \left(\theta_i'', \theta_{-i}\left[\theta_i, \theta_i'\right]\right)\right) \end{aligned} \tag{43}$$

for all θ_i''. Setting

$$z = h\left(\left[\xi_i^1\right]^{-1}\left(\theta_i', \theta_i'\right), \left(\left(\left[\xi_j^1\right]^{-1}\left(\left(\theta_{ij}\left[\theta_i, \theta_i'\right], \theta_{ij}\left[\theta_i, \theta_i'\right]\right)\right)\right)_{j\neq i}\right)\right),$$

condition (42) becomes

$$\begin{aligned}&\sum_{\theta_{-i}\in\Theta_{-i}} u_i\left(z, \left(\theta_i, \theta_{-i}\right)\right)\psi_i\left(\theta_{-i}\right)\left[\theta_i, \theta_i'\right]\\ >\ &\sum_{\theta_{-i}\in\Theta_{-i}} u_i\left(f\left(\theta_i', \theta_{-i}\left[\theta_i, \theta_i'\right]\right), \left(\theta_i, \theta_{-i}\right)\right)\psi_i\left(\theta_{-i}\right)\left[\theta_i, \theta_i'\right].\end{aligned}$$

while condition (43) requires $z \in Y_i\left(\theta_{-i}\left[\theta_i, \theta_i'\right]\right)$. But these latter claims contradict our initial assumption that robust monotonicity fails (i.e., (38)). Thus Bayesian monotonicity fails for this type space and the claim is proved.

($\Leftarrow$) Suppose f satisfies robust monotonicity. Fix any type space $\mathcal{T}$ and any deception α with $f\left(\widehat{\theta}\left(t\right)\right) \neq f\left(\widehat{\theta}\left(\alpha\left(t\right)\right)\right)$ for some t. Define $\boldsymbol{\beta}$ by

$$\beta_i\left(\theta_i\right) = \left\{\theta_i' : \exists t_i \text{ such that } \widehat{\theta}_i\left(t_i\right) = \theta_i \text{ and } \widehat{\theta}_i\left(\alpha_i\left(t_i\right)\right) = \theta_i'\right\}.$$

Deception $\boldsymbol{\beta}$ is unacceptable, so by robust monotonicity, there exist i, θ_i, $\theta_i' \in \beta_i\left(\theta_i\right)$ such that for every $\theta_{-i}' \in \Theta_{-i}$ and $\psi_i \in \Delta\left(\beta_{-i}^{-1}\left(\theta_{-i}'\right)\right)$, there exists $y\left[\theta_{-i}', \psi_i\right] \in Y_i\left(\theta_{-i}'\right)$ such that

$$\sum_{\theta_{-i}\in\Theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(y\left[\theta_{-i}', \psi_i\right], \left(\theta_i, \theta_{-i}\right)\right) > \sum_{\theta_{-i}\in\Theta_{-i}} \psi_i\left(\theta_{-i}\right) u_i\left(f\left(\theta_i', \theta_{-i}'\right), \left(\theta_i, \theta_{-i}\right)\right). \tag{44}$$

Now choose any t_i such that $\widehat{\theta}_i\left(t_i\right) = \theta_i$ and $\widehat{\theta}_i\left(\alpha_i\left(t_i\right)\right) = \theta_i'$. For every (mis-)report θ_{-i}', we now derive a distribution over payoff types θ_{-i} which represents the likelihood that the report θ_{-i}' comes from the true payoff type profile θ_{-i}, given the type space $\mathcal{T}$. For each θ_{-i}', define $\psi_i\left[\theta_{-i}'\right] \in \Delta\left(\Theta_{-i}\right)$ by

$$\psi_i\left(\theta_{-i}\right)\left[\theta_{-i}'\right] \triangleq \frac{\sum_{\left\{t_{-i}:\widehat{\theta}_j\left(\alpha_j\left(t_j\right)\right)=\theta_j' \text{ and } \widehat{\theta}_j\left(t_j\right)=\theta_j,\ \forall j\neq i\right\}} \widehat{\pi}_i\left(t_{-i}\right)\left[t_i\right]}{\sum_{\left\{t_{-i}:\widehat{\theta}_j\left(\alpha_j\left(t_j\right)\right)=\theta_j',\ \forall j\neq i\right\}} \widehat{\pi}_i\left(t_{-i}\right)\left[t_i\right]}. \tag{45}$$

Now let h satisfy

$$h\left(t_i', t_{-i}\right) \triangleq \begin{cases} y\left[\widehat{\theta}_{-i}\left(t_{-i}\right), \psi_i\left[\widehat{\theta}_{-i}\left(t_{-i}\right)\right]\right] & \text{if} \quad t_i' = \alpha_i\left(t_i\right) \\ f\left(\widehat{\theta}\left(t_i', t_{-i}\right)\right) & \text{if} \quad \text{otherwise} \end{cases}. \tag{46}$$

To establish Bayesian monotonicity, it is enough to show that the two inequalities of Bayesian monotonicity are satisfied, or:

$$\sum_{t_{-i}} u_i\left(h\left(\alpha\left(t\right)\right), \widehat{\theta}\left(t\right)\right)\widehat{\pi}_i\left(t_{-i}\right)\left[t_i\right] > \sum_{t_{-i}} u_i\left(f\left(\widehat{\theta}\left(\alpha\left(t\right)\right)\right), \widehat{\theta}\left(t\right)\right)\widehat{\pi}_i\left(t_{-i}\right)\left[t_i\right], \tag{47}$$

and

$$\begin{aligned} & \sum_{t_{-i}} u_i \left(f\left(\widehat{\theta}\left(t_i', t_{-i} \right) \right), \widehat{\theta}\left(t_i', t_{-i} \right) \right) \widehat{\pi}_i \left(t_{-i} \right) \left[t_i' \right] \\ \geq & \sum_{t_{-i}} u_i \left(h\left(\alpha_i \left(t_i \right), t_{-i} \right), \widehat{\theta}\left(t_i', t_{-i} \right) \right) \widehat{\pi}_i \left(t_{-i} \right) \left[t_i \right], \ \forall t_i'. \end{aligned} \tag{48}$$

By inserting the posterior beliefs ψ_i and the rewards $h\left(t_i', t_{-i}\right)$, as defined above in (45) and (46) respectively, we can rewrite the two sides of the inequality (47) as follows:

$$\begin{aligned} & \sum_{t_{-i}} u_i \left(h\left(\alpha\left(t \right) \right), \widehat{\theta}\left(t \right) \right) \widehat{\pi}_i \left(t_{-i} \right) \left[t_i \right] \\ = & \sum_{\theta_{-i}'} \left(\sum_{\left\{ t_{-i} : \widehat{\theta}_j(\alpha_j(t_j)) = \theta_j', \, \forall j \neq i \right\}} \widehat{\pi}_i \left(t_{-i} \right) \left[t_i \right] \right) \sum_{\theta_{-i}} \psi_i \left(\theta_{-i} \right) \left[\theta_{-i}' \right] u_i \left(y\left[\theta_{-i}', \psi_i \left[\theta_{-i}' \right] \right], \theta \right) \end{aligned}$$

and

$$\begin{aligned} & \sum_{t_{-i}} u_i \left(f\left(\widehat{\theta}\left(\alpha\left(t \right) \right) \right), \widehat{\theta}\left(t \right) \right) \widehat{\pi}_i \left(t_i \right) \left[t_{-i} \right] \\ = & \sum_{\theta_{-i}'} \left(\sum_{\left\{ t_{-i} : \widehat{\theta}_j(\alpha_j(t_j)) = \theta_j', \, \forall j \neq i \right\}} \widehat{\pi}_i \left(t_{-i} \right) \left[t_i \right] \right) \sum_{\theta_{-i}} \psi_i \left(\theta_{-i} \right) \left[\theta_{-i}' \right] u_i \left(f\left(\theta' \right), \theta \right) \end{aligned}$$

so (47) follows from (44). Also

$$\begin{aligned} & \sum_{t_{-i}} u_i \left(h\left(\alpha_i \left(t_i \right), t_{-i} \right), \widehat{\theta}\left(t_i', t_{-i} \right) \right) \widehat{\pi}_i \left(t_{-i} \right) \left[t_i' \right] \\ = & \begin{cases} \sum_{t_{-i}} u_i \left(y\left[\widehat{\theta}_{-i}\left(t_{-i} \right), \psi_i \left[\widehat{\theta}_{-i}\left(t_{-i} \right) \right] \right], \widehat{\theta}\left(t_i', t_{-i} \right) \right) \widehat{\pi}_i \left(t_{-i} \right) \left[t_i' \right] & \text{if} \quad t_i' = \alpha_i \left(t_i \right) \\ \sum_{t_{-i}} u_i \left(f\left(\widehat{\theta}\left(t_i', t_{-i} \right) \right), \widehat{\theta}\left(t_i', t_{-i} \right) \right) \widehat{\pi}_i \left(t_{-i} \right) \left[t_i' \right], & \text{if} \quad t_i' \neq \alpha_i \left(t_i \right) \end{cases} \end{aligned}$$

Now $y\left[\widehat{\theta}_{-i}\left(t_{-i} \right), \psi_i \left[\widehat{\theta}_{-i}\left(t_{-i} \right) \right] \right] \in Y_i \left(\widehat{\theta}_{-i}\left(t_{-i} \right) \right)$ implies (48).

The proof may appear rather intricate in its details. We next give a brief outline of the basic steps to show how interim implies robust monotonicity. The proof proceeds by contrapositive. We start with an unacceptable deception β which by hypothesis fails robust monotonicity and hence satisfies the inequalities (38) and (39). For the given deception β, we then create a type space, consisting of two components for every agent i. The first component for agent i is created by the set of pairs of payoff types $\left(\theta_i, \theta_i' \right)$, where the first entry is the true payoff type and the second entry is a feasible deception (under β), or $\theta_i' \in \beta_i \left(\theta_i \right)$. For this reason, we refer to these types as "deception types." For every such pair $\left(\theta_i, \theta_i' \right)$ there exists at least one particular payoff profile θ_{-i}'

which acts as a misreport. Under the deception β, this payoff profile θ'_{-i} could have been reported by all true payoff profiles which are in the support of ψ_i. Consequently, the belief component of type $\left(\theta_i, \theta'_i\right)$ is given by simply adopting $\psi_i\left(\cdot \left| \theta_i, \theta'_i\right.\right)$. The second component consists of "pseudo complete information types", described by $t_i = \theta \in \Theta$. Each such type has a belief that assigns probability one to the event that the true payoff profile is given by θ and that all other agents report the deception type (θ_j, θ_j), and hence the "pseudo" in the labelling.

Given this type space T_i, we then consider a particular deception $\alpha_i : T_i \to T_i$. The deception α_i is localized around the "deception types" and the "pseudo complete information types" report thruthfully. The deception α_i consists of agent i always reporting his deception type rather than his true type, or $\alpha_i\left(\theta_i, \theta'_i\right) = \left(\theta'_i, \theta'_i\right)$. We then verify whether f is interim monotone under α. The existence of the pseudo complete information types θ forces the interim incentive compatibility conditions to reduce to ex post incentive compatibility conditions. This guarantees the hypothesis in the robust monotonicity notion, namely inequality (38), and thus leads to the conclusion in form of the inequalities (39). But then we obtain a contradiction to the reward condition of interim monotonicity, unless the hypothesis for the interim monotonicity condition, namely $f \neq f \circ \alpha$, is not satisfied, i.e. $f = f \circ \alpha$ holds, but of course this implies that β is acceptable.

References

Abreu, D., and H. Matsushima (1992a): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60, 993–1008.

——— (1992b): "Virtual Implementation In Iteratively Undominated Strategies: Incomplete Information," Discussion paper, Princeton University and University of Tokyo.

Battigalli, P. (1999): "Rationalizability in Incomplete Information Games," Discussion Paper ECO 99/17, European University Institute.

Battigalli, P., and M. Siniscalchi (2003): "Rationalization and Incomplete Information," *Advances in Theoretical Economics*, 3, Article 3.

Benoit, J., and E. Ok (2008): "Nash Implementation Without No Veto," *Games and Economic Behavior*, forthcoming.

Bergemann, D., and S. Morris (2005a): "Robust Implementation: The Role of Large Type Spaces," Discussion Paper 1519, Cowles Foundation, Yale University.

——— (2005b): "Robust Mechanism Design," *Econometrica*, 73, 1771–1813.

——— (2007): "Robust Implementation in Direct Mechanisms," Discussion Paper 1561R, Cowles Foundation, Yale University.

——— (2008a): "Ex Post Implementation," *Games and Economic Behavior*, 63, 527–566.

——— (2008b): "Robust Implementation in General Mechanisms," Discussion paper, Cowles Foundation, Yale University.

——— (2008c): "Strategic Distinguishability and Robust Virtual Implementation," Discussion Paper 1609R, Cowles Foundation, Yale University.

Borgers, T. (1995): "A Note on Implementation and Strong Dominance," in *Social Choice, Welfare and Ethics*, ed. by W. Barnett, H. Moulin, M. Salles, and N. Schofield. Cambridge University Press, Cambridge.

Brandenburger, A., and E. Dekel (1987): "Rationalizability and Correlated Equilibria," *Econometrica*, 55, 1391–1402.

Chung, K.-S., and J. Ely (2001): "Efficient and Dominance Solvable Auctions with Interdependent Valuations," Discussion paper, Northwestern University.

Dekel, E., D. Fudenberg, and S. Morris (2007): "Interim Correlated Rationalizability," *Theoretical Economics*, 2, 15–40.

Duggan, J. (1997): "Virtual Bayesian Implementation," *Econometrica*, 65, 1175–1199.

Gale, D. (1989): *The Theory of Linear Economic Models*. University of Chicago Press, Chicago.

Hughes, G., and M. Creswell (1996): *A New Introduction Into Modal Logic*. Routledge, London.

Jackson, M. (1991): "Bayesian Implementation," *Econometrica*, 59, 461–477.

Jackson, M. (1992): "Implementation in Undominated Strategies: A Look at Bounded Mechanisms," *Review of Economic Studies*, 59, 757–775.

Lipman, B. (1994): "A Note on the Implications of Common Knowledge of Rationality," *Games and Economic Behavior*, 6, 114–129.

Palfrey, T., and S. Srivastava (1989): "Mechanism Design with Incomplete Information: A Solution to the Implementation Problem," *Journal of Political Economy*, 97, 668–691.

Postlewaite, A., and D. Schmeidler (1986): "Implementation in Differential Information Economies," *Journal of Economic Theory*, 39, 14–33.

Serrano, R., and R. Vohra (2001): "Some Limitations of Virtual Bayesian Implementation," *Econometrica*, 69, 785–792.

——— (2005): "A Characterization of Virtual Bayesian Implementation," *Games and Economic Behavior*, 50, 312–331.

Wilson, R. (1987): "Game-Theoretic Analyses of Trading Processes," in *Advances in Economic Theory: Fifth World Congress*, ed. by T. Bewley, pp. 33–70, Cambridge. Cambridge University Press.