Robust Virtual Implementation with Incomplete Information: Towards a Reinterpretation of the Wilson Doctrine

Georgy Artemov*
Takashi Kunimoto†
and
Roberto Serrano‡

This version: May 2007

Abstract

We consider robust virtual implementation, where robustness is the requirement that implementation succeed in all type spaces consistent with a given payoff type space as well as with a given space of first-order beliefs about the other agents' payoff types. This last bit, which constitutes our reinterpretation of the Wilson doctrine, allows us to obtain very permissive results. Our first result is that generically, if there are at least three alternatives, any incentive compatible social choice function is robustly virtually implementable in iteratively undominated strategies. Further, we characterize robust virtual implementation in iteratively undominated strategies by means of incentive compatibility and measurability. Our characterization is independent of the presence of monetary transfers or assumptions alike, made in previous studies. Our work also clarifies the measurability condition in connection to the generic diversity of preferences used in our first result.

JEL Classification: C72, D78, D82.

Keywords: Wilson doctrine, mechanism design, robust virtual implementation, iteratively undominated strategies, incentive compatibility, measurability, type diversity.

*Department of Economics, Brown University, Providence, RI, U.S.A.; ga@brown.edu

†Department of Economics, McGill University, Montreal, Canada; takashi.kunimoto@mcgill.ca

‡Department of Economics, Brown University, Providence, RI, U.S.A. and IMDEA-Ciencias Sociales, Madrid, Spain; roberto_serrano@brown.edu

"Game Theory has a great advantage in explicitly analyzing the consequences of trading rules that presumably are really common knowledge; it is deficient to the extent it assumes other features to be common knowledge, such as one player's probability assessment about another's preferences or information.
I foresee the progress of game theory as depending on successive reductions in the base of common knowledge required to conduct useful analysis of practical problems. Only by repeated weakening of common knowledge assumptions will the theory approximate reality."

Robert Wilson (1987)

1 Introduction

The theory of implementation attempts to identify the conditions under which a social choice rule may be decentralized; that is, when agents, acting on their self-interest, arrive at the outcomes prescribed by the social choice rule. In contexts in which the economic authority knows what agents' types might be, but does not know what they actually are, the theory has uncovered necessary and sufficient conditions for such decentralization.[1] In many circumstances, one should expect that, apart from the economic authority's informational constraints, agents themselves be also asymmetrically informed about each other's preferences, beliefs or signals.

For such *incomplete information environments*, a necessary condition for the implementation of any rule is its *incentive compatibility*. Some authors refer to this condition as informational feasibility, and give it the same stature as physical feasibility (e.g., Myerson (1989)): by the revelation principle, a rule is *truthfully implementable* in Bayesian equilibrium if and only if it is incentive compatible. Yet the direct revelation mechanism that yields a truthfully implementable rule will typically have additional equilibria, and these equilibria are undesirable in the sense of not being consistent with the original social choice rule. This motivates the question of *full implementation*: the search for mechanisms whose entire set of equilibrium outcomes relates to the given rule. Full implementation in incomplete information

[1] See Jackson (2001), Maskin and Sjostrom (2002), Palfrey (2002) or Serrano (2004) for recent surveys.

environments will be the notion of implementation sought in the current paper.

When the set of equilibrium outcomes is required to coincide with those picked out by the rule, we speak of *exact implementation*. A new necessary condition – *Bayesian monotonicity* – emerges in this case in addition to incentive compatibility (Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989), Jackson (1991)). Moreover, Jackson (1991) finds the version of this condition that, along with incentive compatibility and other assumptions, yields a characterization of Bayesian implementable rules.

It is well-known that Bayesian monotonicity may sometimes be a very restrictive condition (e.g., Palfrey and Srivastava (1987), Chakravorti (1992)). In view of this, one can relax the requirement of exact implementation, and, instead, ask that the set of equilibrium outcomes approximates the rule. This is the approach known as *virtual implementation*, which has confined its scope to social choice functions (SCFs). Though some new sufficient conditions accompanying incentive compatibility were identified (*incentive consistency* in Duggan (1997), *measurability* in Abreu and Matsushima (1992)), they were not necessary conditions, and not even logically weaker than Bayesian monotonicity (as shown in Serrano and Vohra (2001)). Finally, Serrano and Vohra (2005) identify the condition of *virtual monotonicity*, which, along with incentive compatibility, characterizes virtual implementation in Bayesian equilibrium. It is argued there that virtual monotonicity is an extremely weak condition, strictly weaker than Bayesian monotonicity and measurability, and trivially satisfied by all SCFs in "most" environments.

From the view-point of the realism of the approach, all these papers have an important drawback. Following the Wilson doctrine, expressed in the quote by Wilson (1987) at the beginning of our introduction, the theory should aim to relax undesirable common knowledge assumptions among the agents. In particular, one should avoid in mechanisms the use of the notion of a *type*. A type, which includes the specification of higher-order beliefs for a player, may well be far too complex an object to describe. Accepting this view, the usual route taken by researchers has been to prevent the use of any consideration of beliefs in the message spaces. Thus, mechanisms have been constructed on the basis of only that part of the type that is payoff relevant, the so-called *payoff type*.

In a series of papers, Bergemann and Morris (2005a, 2005b, 2007) seek for robust

implementation results. Their work relevant to the current paper is contained in their latter two papers, which deal with full implementation. Faithful to the Wilson doctrine, they construct mechanisms that rely exclusively on the use of payoff types, and require that implementation must obtain for any type space consistent with the original payoff type space. When insisting on *robust exact implementation*, Bergemann and Morris identify *ex-post incentive compatibility* and *robust monotonicity* as necessary and almost sufficient conditions. They also consider *robust virtual implementation* and identify *ex-post incentive compatibility* and *robust measurability* as the corresponding key conditions for this case. These conditions are very restrictive, stronger than their counterparts for exact or virtual implementation for a fixed type space. For instance, ex-post incentive compatibility would generically require an SCF to be constant (Jehiel et al. (2006)). Bergemann and Morris interpret their negative results for virtual implementation as a consequence of the robustness approach. They suggest that "the distance between [exact] and virtual implementation may shrink considerably after imposing robustness on the implementation concept" (Bergemann and Morris, 2005b, p. 42) and that the constructions for virtual implementation rely "on the implicit assumption that there is a common knowledge of mapping from beliefs to payoff types of all agents (a "Beliefs-Determine-Preferences" property)" (Bergemann and Morris, 2007, p. 5). We shall offer now a different interpretation.

To show the difficulties of robust virtual implementation, Bergemann and Morris (2005b, 2007) construct a very specific type space in which the interim preferences of all types are aligned. That virtual Bayesian implementation may fail exactly for this reason has already been pointed out in Serrano and Vohra (2001) in a standard Bayesian environment with a fixed type space. Yet, such failures are "rare:" if the environment satisfies *type diversity*, every incentive compatible SCF is virtually implementable (Serrano and Vohra (2005)). That is, virtual implementation is as successful as it can possibly be. Furthermore, in environments with at least three alternatives, type diversity is *generically* satisfied.

This paper reconsiders the problem of robust virtual implementation. Our approach is the following. First, we take from the foundations of game theory that, when requiring robustness results with respect to consistent type spaces, equilibrium restrictions are not imposed beyond the ones identified by the weaker solution concept of iterative elimination of strictly dominated strategies. This will be the solution

concept we shall employ, and in doing so, we are building on an important paper by Abreu and Matsushima (1992).[2]

Our main point of departure from the usual interpretation of the Wilson doctrine is that we shall allow the use of first-order beliefs over payoff types, along with payoff types, in our mechanisms. This is a reinterpretation of the Wilson doctrine that does not seem too demanding: after all, people are usually capable of providing simple probability assessments; we could think here of an insurance problem, for instance. The combination of payoff type and first-order belief for a player will comprise our notion of a *pseudo-type*. Therefore, we shall fix a (typically quite rich) space of pseudo-types, which we will assume to be common knowledge among the agents, and we shall require that implementation obtains for all type spaces consistent with our original pseudo-type space.

The resulting theorems we obtain send a very different message from the one sent by Bergemann and Morris (2005b, 2007) for robust virtual implementation. First, we propose a condition that we term *pseudo-type diversity* and show (Theorem 1) that in such environments every SCF that is *incentive compatible in every type space consistent with the original pseudo-type space* is robustly virtually implementable in iteratively undominated strategies. Thus, there is no need to go all the way to requiring ex-post incentive compatibility – the relevant interim notion for every pseudo-type in the model suffices. Second, pseudo-type diversity again happens to be generic when there are at least three alternatives; thus, one does not need to rely on any additional condition "most of the time."

Next, we seek to obtain a characterization. We extend the work of Abreu and Matsushima (1992) to our settings. Theorem 2 shows that *incentive compatibility for every consistent type space* and *A-M measurability* – introduced in Abreu and Matsushima (1992) – are necessary and sufficient conditions for robust virtual

[2]Following Bergemann and Morris (2005b) and Brandenburger and Dekel (1987), we can also characterize our solution concept – iteratively undominated strategies – in terms of interim rationalizability which, in turn, is equivalent to the Bayesian equilibria in all consistent type spaces. There are, however, two reasons why our definition of interim rationalizability is more demanding than that of Bergemann and Morris (2005b). First, we include the set of first-order beliefs over the payoff type space as part of the environment which is assumed to be common knowledge. Second, at each round of elimination of never best responses, we explicitly require agents' first-order beliefs to be consistent with the environment. This is termed Δ-rationalizability in Battigalli and Siniscalchi (2003).

implementation in iteratively undominated strategies. Importantly, we relax an assumption made by Abreu and Matsushima on the environment, which essentially amounts to quasilinear utilities on a nummeraire, on which small punishments are imposed. Moreover, we elaborate on the connection between pseudo-type diversity and A-M measurability: as hinted in the original paper by Abreu and Matsushima, the pseudo-type diversity condition is associated with the first iteration of the measurability algorithm, which, in general, may have multiple steps. The algorithm determines the maximum possible separation of types – or pseudo-types – on the basis of their interim preferences.[3] We also note that the proofs of our Theorems 1 and 2 follow the same logic, further underscoring the link between pseudo-type diversity and measurability.

A final word is called for regarding the nature of our mechanisms and the connection with virtual implementation in Bayesian equilibrium. First, the distinction between implementation in pure- or mixed-strategy equilibria is of no significance, once we ask for robustness with respect to type spaces. Our sufficiency result applies a fortiori to virtual implementation in mixed-strategy Bayesian equilibrium. Although virtual implementation in Bayesian equilibrium is typically more permissive than virtual implementation in iteratively undominated strategies,[4] the difference is small in that it concerns environments violating pseudo-type diversity. Furthermore, the additional SCFs so implemented must rely on the use of non-regular mechanisms (e.g., using integer games and devices alike): following a result of Abreu and Matsushima (1992) for a fixed type space, A-M measurability is necessary for robust virtual implementation in Bayesian equilibrium if one uses regular mechanisms. Our mechanisms are finite, and best responses always exist.

The paper is organized as follows: in Section 2 we introduce the preliminary notation and definitions. In Section 3 we present our first mechanism, which is used in Section 4 to prove our main result (Theorem 1). Section 5 is concerned with A-M measurability, used in Section 6 for the characterization result (Theorem 2). In Section 7 we explain the connection of our results with those in virtual Bayesian implementation. We conclude in Section 8.

[3]See also a related discussion of indistinguishability in Bergemann and Morris (2007).

[4]For each fixed type space, this follows since virtual monotonicity is strictly weaker than A-M measurability.

2 Preliminaries

Let $N = \{1, \ldots, n\}$ denote the set of agents and Θ_i be the set of finite payoff-relevant (or, simply, *payoff*) types of agent i. Denote $\Theta \equiv \Theta_1 \times \cdots \times \Theta_n$, and $\Theta_{-i} \equiv \Theta_1 \times \cdots \times \Theta_{i-1} \times \Theta_{i+1} \times \cdots \Theta_n$.[5] Let $q_i(\theta_{-i}|\theta_i)$ denote agent i's first-order belief that other agents receive the profile of payoff types θ_{-i} when his payoff type is θ_i. Let Q_i be the finite set of such probabilistic first-order beliefs of agent i. We call $T_i = \Theta_i \times Q_i$ the finite set of *pseudo-types* of agent i. Agent i's pseudo-type t_i contains information about his payoff type θ_i and the first-order belief over Θ_{-i} conditional on θ_i.

Let A denote the set of pure outcomes, which are assumed to be independent of the information state. Let $\mathcal{A}$ be a σ-algebra on A and $\Delta(A)$ denote the set of probability measures on $(A, \mathcal{A})$ with countable supports.

Agent i's state dependent von Neumann-Morgenstern utility function is denoted $u_i : \Delta(A) \times \Theta \to \mathbb{R}$.

We can now define an *environment* as $\mathcal{E} = ((A, \mathcal{A}), \{u_i, \Theta_i, Q_i\}_{i \in N})$, which is implicitly understood to be common knowledge among the agents.

We denote a type of agent i by τ_i and the agent i's set of types by $\mathcal{T}_i$. A *type* τ_i of agent i must include a description of his pseudo-type, which in turn includes a payoff type. Thus, there is a function $\hat{t}_i : \mathcal{T}_i \to T_i$, with $\hat{t}_i(\tau_i)$ being agent i's pseudo-type when his type is τ_i. We shall write $\hat{t}(\tau)$ to refer to the profile of pseudo-types when the type profile is τ. There is also a function $\hat{\theta}_i : \mathcal{T}_i \to \Theta_i$, with $\hat{\theta}_i(\tau_i)$ being agent i's payoff type when his type is τ_i. We shall write $\hat{\theta}(\tau)$ to denote the payoff type profile when the profile of types is τ. With some abuse of notation, let $\hat{\theta}_i(t_i)$ be agent i 's payoff type when his pseudo-type is t_i. A type τ_i of agent i must also include a description of his beliefs about the types of the other agents; thus, for any $\tau_{-i} \in \mathcal{T}_{-i}$, $\pi_i(\tau_{-i}|\tau_i)$ denotes the probability that agent i of type τ_i assigns to other agents having types τ_{-i}.

We require that types, pseudo-types and payoff types are consistent with each other. We express the consistency requirement in the following definition. A *type space* $\mathcal{T}$ is a collection:

$$\mathcal{T} = (\mathcal{T}_i, \hat{\theta}_i, \hat{t}_i, \pi_i)_{i \in N}.$$

[5] Similar notation will be used for products of other sets.

Definition 1 *A type space* $\mathcal{T} \equiv (\mathcal{T}_i, \hat{\theta}_i, \hat{t}_i, \pi_i)_{i \in N}$ *is said to be* ***consistent*** *with an environment* $\mathcal{E} = ((A, \mathcal{A}), \{u_i, \Theta_i, Q_i\}_{i \in N})$ *if, for every* $i \in N$ *and every type* $\tau_i \in \mathcal{T}_i$*, the following two conditions must hold:*

1. $\hat{\theta}_i(\tau_i) \in \Theta_i$ *and* $\hat{t}_i(\tau_i) \in \Theta_i \times Q_i$*; and*

2. $\hat{\theta}_i(\tau_i) = \theta_i$ *whenever* $\hat{t}_i(\tau_i) = (\theta_i, q_i)$ *for some* $(\theta_i, q_i) \in \Theta_i \times Q_i$*.*

The first part of the definition is just the requirement that pseudo-type and payoff type be consistent with the agent's type. This requirement, for payoff types, has also been imposed in Bergemann and Morris (2005a, 2005b, 2007). The second part requires similar consistency between pseudo-types and payoff types. These two requirements, in turn, imply that, for any $\tau_i \in \mathcal{T}_i$ with $\hat{t}_i(\tau_i) = t_i = (\hat{\theta}_i(t_i), q_i)$,

$$\sum_{\tau_{-i}: \hat{t}_{-i}(\tau_{-i}) = t_{-i}} \pi_i(\tau_{-i} | \tau_i) = q_i(\hat{\theta}_{-i}(t_{-i}) | \hat{\theta}_i(t_i))$$

The consistency we have just defined essentially reduces to the requirement that the various levels of beliefs of an individual do not contradict one another. This requirement is the same as common knowledge of coherency, which is imposed when Brandenburger and Dekel (1993) and Mertens and Zamir (1985) construct the universal type space. The only difference here is that the underlying state space – the pseudo-type space – includes not only the payoff type space but also the set of the first-order beliefs over the payoff type space.

When a consistent type space $\mathcal{T}$ satisfies the properties that $\mathcal{T}_i = \Theta_i$ and Q_i is a singleton for each agent $i \in N$, then the true type space is common knowledge. This corresponds to the fixed Bayesian environment (e.g., Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989), Jackson (1991), Abreu and Matsushima (1992), Duggan (1997), Serrano and Vohra (2001, 2005)). When Q_i includes any possible belief of i over Θ_{-i} – that is, Q_i is not assumed to be common knowledge among agents in N – this corresponds to the payoff environment of Bergemann and Morris (2005b, 2007). Our approach is in between these two extremes, as it allows Q_i to include an arbitrarily large, but finite, number of beliefs. We note that Bergemann and Morris (2007) also make the finiteness assumption on the space of payoff types.

A *social choice function* (SCF) is a function $f : T \to \Delta(A)$. Note that the domain of the SCFs is not the true type space, but the pseudo-type space. That is, the goals of society do not depend on the possibly far too complex higher-order beliefs structure, but they might depend on payoff relevant information, as well as on first-order beliefs about it.[6]

Fix any consistent type space $\mathcal{T}$ throughout. The interim expected utility of agent i of type τ_i corresponding to an SCF f is defined as:

$$U_i(f|\tau_i) \equiv \sum_{\tau_{-i} \in \mathcal{T}_{-i}} \pi_i(\tau_{-i}|\tau_i) u_i(f(\hat{t}(\tau_i, \tau_{-i})); \hat{\theta}(\tau_i, \tau_{-i}))$$

A *mechanism* $\Gamma = ((M_i)_{i \in N}, g)$ describes a message space M_i for agent i and an outcome function $g : M \to \Delta(A)$, where $M = \times_{i \in N} M_i$. Let $\sigma_i : \mathcal{T}_i \to M_i$ denote a (pure) strategy for agent i and Σ_i his set of pure strategies. Let

$$U_i(g \circ \sigma|\tau_i) \equiv \sum_{\tau_{-i} \in \mathcal{T}_{-i}} \pi_i(\tau_{-i}|\tau_i) u_i(g(\sigma(\tau_{-i}, \tau_i)); \hat{\theta}(\tau_{-i}, \tau_i)).$$

Given a mechanism $\Gamma = (M, g)$, let H_i be a subset of Σ_i.

Definition 2 (Strict Dominance) [7] *A strategy $\sigma_i \in H_i$ is **strictly dominated** for player i with respect to $H = \times_{j \in N} H_j$ if there exist $\tau_i \in \mathcal{T}_i$ and $\sigma_i' \in H_i$ such that for every $\sigma_{-i} \in \times_{j \neq i} H_j$,*

$$U_i(g \circ (\sigma_i', \sigma_{-i})|\tau_i) > U_i(g \circ (\sigma_i, \sigma_{-i})|\tau_i).$$

[6] This last bit allows our model to cover environments in which the only uncertainty concerns information, as in basic insurance problems (see, for example, Serrano and Vohra (2001, Example 1)). To accomodate that example into our model, since no uncertainty exists over payoff types, one can add a space of signals over which first-order beliefs are defined.

[7] We use the same definition of strict dominance as Abreu and Matsushima (1992), yet we note that we could obtain our results with the less demanding notion of dominance, which require a strategy to be dominated for each type τ_i:
Definition: A strategy $\sigma_i \in H_i$ is **strictly dominated** for agent i with respect to $H = \times_{j \in N} H_j$ if for each $\tau_i \in \mathcal{T}_i$ there exists $\sigma_i' \in H_i$ such that for every $\sigma_{-i} \in \times_{j \neq i} H_j$,

$$U_i(g \circ (\sigma_i', \sigma_{-i})|\tau_i) > U_i(g \circ (\sigma_i, \sigma_{-i})|\tau_i).$$

Let $\mathcal{K}_i(H)$ denote the set of all undominated strategies for agent i with respect to $H = \times_{i\in N} H_i$. Let $\mathcal{K}(H) = \times_{i\in N}\mathcal{K}_i(H)$. Let $\mathcal{K}_i^0(\Sigma) = \Sigma_i$ and for each $k \geq 1$, $\mathcal{K}^k(\Sigma) = \times_{i\in N}\mathcal{K}_i^k(\Sigma)$, where $\Sigma = \times_{i\in N}\Sigma_i$ and $\mathcal{K}_i^k(\Sigma) = \mathcal{K}_i(\mathcal{K}^{k-1}(\Sigma))$. Let

$$\mathcal{K}^* \equiv \bigcap_{k=0}^{\infty} \mathcal{K}^k(\Sigma)$$

Definition 3 (Iterative Dominance) *A strategy profile $\sigma \in \Sigma$ is **iteratively undominated** if $\sigma \in \mathcal{K}^*$.*

An SCF f is said to be *exactly implementable* in iteratively undominated strategies for a type space $\mathcal{T}$ if there exists a mechanism $\Gamma = (M, g)$ such that for any $\sigma \in \mathcal{K}^*$, $g(\sigma(\tau)) = f(\hat{t}(\tau))$ for all $\tau \in \mathcal{T}$. We add the requirement that this definition should hold for every consistent type space $\mathcal{T}$ to obtain the definition of robust implementation:

Definition 4 (Robust Implementation) *An SCF f is said to be **robustly implementable** in iteratively undominated strategies if there exists a mechanism $\Gamma = (M, g)$ such that for any $\sigma \in \mathcal{K}^*$, $g(\sigma(\tau)) = f(\hat{t}(\tau))$ for every $\tau \in \mathcal{T}$ and every consistent type space $\mathcal{T}$.*

Consider the following metric on SCFs:

$$d(f, h) = \sup\left\{|f(t|S) - h(t|S)| \;\middle|\; t \in T,\ S \in \mathcal{A}\right\}$$

The notation $f(t|S)$ refers to the lottery $f(t) \in \Delta(A)$ when its support is restricted to $S \in \mathcal{A}$.

An SCF f is said to be *virtually implementable* in iteratively undominated strategies for a consistent type space $\mathcal{T}$ if, there exists $\bar{\varepsilon} > 0$ such that for any $\varepsilon \in (0, \bar{\varepsilon}]$, there exists an SCF f^ε for which $d(f, f^\varepsilon) < \varepsilon$ and f^ε is exactly implementable in iteratively undominated strategies for the type space $\mathcal{T}$.

The definition of implementability that will be used in this paper follows:

Definition 5 (Robust Virtual Implementation) *An SCF f is **robustly virtually implementable** in iteratively undominated strategies if there exists $\bar{\varepsilon} > 0$ such*

that, for any $\varepsilon \in (0, \bar{\varepsilon}]$*, there exists an SCF* f^{ε} *for which* $d(f, f^{\varepsilon}) < \varepsilon$ *and* f^{ε} *is robustly implementable in iteratively undominated strategies.*

The next standard definition is very important in the entire economic theory of information:

Definition 6 (Incentive Compatibility) *An SCF* $f : T \to \Delta(A)$ *is said to be* ***incentive compatible*** *for a consistent type space* $\mathcal{T}$ *if for every* $i \in N,\ \tau_i, \tau_i^{'} \in \mathcal{T}_i$*,*

$$\sum_{\tau_{-i} \in \mathcal{T}_{-i}} \pi_i(\tau_{-i}|\tau_i) u_i(f(\hat{t}(\tau_i, \tau_{-i})); \hat{\theta}(\tau_i, \tau_{-i})) \geq \sum_{\tau_{-i} \in \mathcal{T}_{-i}} \pi_i(\tau_{-i}|\tau_i) u_i(f(\hat{t}(\tau_i^{'}, \tau_{-i})); \hat{\theta}(\tau_i, \tau_{-i}))$$

We shall say that an SCF f is *strictly* incentive compatible if all the inequalities in the preceding definition are strict whenever $\hat{t}_i(\tau_i) \neq \hat{t}_i(\tau_i^{'})$.

Define $V_i(f|t_i)$ to be the interim expected utility of agent i of pseudo-type t_i corresponding to an SCF f as follows:

$$V_i(f|t_i) = \sum_{\theta_{-i} \in \Theta_{-i}} \sum_{q_{-i} \in Q_{-i}} q_i(\theta_{-i}|\theta_i) u_i(f(t_i, \theta_{-i}, q_{-i}); \theta_i, \theta_{-i})$$

where $t_i \equiv (\theta_i, q_i) \in T_i = \Theta_i \times Q_i$. We call $V_i(f|t_i)$ the *pseudo*-interim utility of agent i.

This notion suggests the following definition:

Definition 7 (Pseudo-Incentive Compatibility) *An SCF* $f : T \to \Delta(A)$ *is* ***pseudo-incentive compatible*** *if, for any* $i \in N$ *and any* $t_i = (\theta_i, q_i), t_i^{'} = (\theta_i^{'}, q_i^{'}) \in T_i$ *with* $t_i \neq t_i^{'}$*,*

$$\begin{aligned} &\sum_{\theta_{-i} \in \Theta_{-i}} \sum_{q_{-i} \in Q_{-i}} q_i(\theta_{-i}|\theta_i) u_i(f(t_i, \theta_{-i}, q_{-i}); \theta_i, \theta_{-i}) \\ \geq &\sum_{\theta_{-i} \in \Theta_{-i}} \sum_{q_{-i} \in Q_{-i}} q_i(\theta_{-i}|\theta_i) u_i(f(t_i^{'}, \theta_{-i}, q_{-i}); \theta_i, \theta_{-i}) \end{aligned}$$

We shall say that an SCF f is *strictly* pseudo-incentive compatible if all the inequalities in the preceding definition are strict.

The next lemma provides a useful link between these concepts:

Lemma 1 *An SCF $f : T \to \Delta(A)$ is incentive compatible for any consistent type space $\mathcal{T}$ if and only if it is pseudo-incentive compatible.*

Proof of Lemma 1: Fix an arbitrary consistent type space $\mathcal{T}$. For each $\tau_i \in \mathcal{T}_i$, let $\hat{t}_i(\tau_i) \equiv t_i$ and $\hat{\theta}_i(\tau_i) \equiv \theta_i$.

Incentive compatibility for the consistent type space $\mathcal{T}$

$$\Leftrightarrow \sum_{\tau_{-i} \in \mathcal{T}_{-i}} \pi_i(\tau_{-i}|\tau_i) \left[u_i(f(\hat{t}(\tau)); \hat{\theta}(\tau)) - u_i(f(\hat{t}(\tau_i', \tau_{-i})); \hat{\theta}(\tau)) \right] \geq 0$$

$$\Leftrightarrow \sum_{t_{-i} \in T_{-i}} \sum_{\tau_{-i}: \hat{t}_{-i}(\tau_{-i}) = t_{-i}} \pi_i(\tau_{-i}|\tau_i) \left[u_i(f(t_i, t_{-i}); \hat{\theta}(t_i, t_{-i})) - u_i(f(t_i', t_{-i})); \hat{\theta}(t)) \right] \geq 0$$

($\because [\cdot]$ is the same for every $\tau_{-i} : \hat{t}_{-i}(\tau_{-i}) = t_{-i}$)

$$\Leftrightarrow \sum_{t_{-i} \in T_{-i}} \left[u_i(f(t_i, t_{-i}); \hat{\theta}(t_i, t_{-i})) - u_i(f(t_i', t_{-i})); \hat{\theta}(t)) \right] \sum_{\tau_{-i}: \hat{t}_{-i}(\tau_{-i}) = t_{-i}} \pi_i(\tau_{-i}|\tau_i) \geq 0$$

$$\Leftrightarrow \sum_{\theta_{-i} \in \Theta_{-i}} \sum_{q_{-i} \in Q_{-i}} q_i(\theta_{-i}|\theta_i) \left[u_i(f(t_i, \theta_{-i}, q_{-i}); \theta_i, \theta_{-i})) - u_i(f(t_i', \theta_{-i}, q_{-i})); \theta_i, \theta_{-i}) \right] \geq 0$$

($\because \ t_{-i} = (\theta_{-i}, q_{-i}) \in \Theta_{-i} \times Q_{-i}$)

$\Leftrightarrow$ pseudo-incentive compatibility.■

As is well-known, the next proposition identifies incentive compatibility as a necessary condition for implementability:

Proposition 1 *If an SCF is robustly virtually implementable in iteratively undominated strategies, then it is incentive compatible for every consistent type space.*

Proof of Proposition 1: By our hypothesis, there exists an SCF f^ε such that $d(f^\varepsilon, f) < \varepsilon$ and f^ε is robustly exactly implementable in iteratively undominated strategies for any consistent type space $\mathcal{T}$. Fix an arbitrary consistent type space $\mathcal{T}$. Suppose that f is not incentive compatible; that is, the weak inequality in definition 6 does not hold. Then, there exists a small enough $\varepsilon > 0$ such that the same inequality for f^ε does not hold either. Therefore, f^ε is not incentive compatible and, thus, cannot be exactly implementable, a contradiction.

Since the same argument holds for any consistent type space, one can conclude that for $\varepsilon > 0$ small enough, f is incentive compatible for every consistent type space if and only if f^ε is incentive compatible for every consistent type space. ■

We shall make the following weak regularity assumption on environments:

Definition 8 (Pseudo-NTI) *An environment $\mathcal{E}$ satisfies* ***pseudo-no-total-indifference (pseudo-NTI)*** *if for every $i \in N$ and $t_i = (\theta_i, q_i) \in T_i$, there exist $a, a' \in A$ such that*

$$\sum_{\theta_{-i} \in \Theta_{-i}} q_i(\theta_{-i}|\theta_i) u_i(a; \theta_i, \theta_{-i}) \neq \sum_{\theta_{-i} \in \Theta_{-i}} q_i(\theta_{-i}|\theta_i) u_i(a'; \theta_i, \theta_{-i}).$$

Pseudo-NTI simply rules out indifference (in terms of pseudo-interim expected utility) across all lotteries.

Let $A = \{a_1, \ldots, a_K\}$ be the finite set of alternatives.[8] Henceforth, we will find it convenient to identify a lottery $x \in \Delta(A)$ as a point in the $(K-1)$ dimensional unit simplex $\Delta^{K-1} = \{(x_1, \ldots, x_K) \in \mathbb{R}_+^{K-1} \mid \sum_{k=1}^{K} x_k = 1\}$. Define $V_i^k(t_i)$ to be the interim expected utility of agent i of pseudo-type $t_i = (\theta_i, q_i)$ for the constant SCF that assigns a_k in each state in T, i.e.,

$$V_i^k(t_i) = \sum_{\theta_{-i} \in \Theta_{-i}} q_i(\theta_{-i}|\theta_i) u_i(a_k; \theta_i, \theta_{-i}).$$

Let $V_i(t_i) = (V_i^1(t_i), \ldots, V_i^K(t_i))$.

Next, we define the condition of pseudo-type diversity in an environment, which will play an important role in our analysis:

Definition 9 (Pseudo-TD) *An environment $\mathcal{E}$ satisfies* ***pseudo-type diversity (pseudo-TD)*** *if there do not exist $i \in N$, $t_i = (\theta_i, q_i), t_i' = (\theta_i', q_i') \in T_i$ with $t_i \neq t_i'$, $\beta \in \mathbb{R}_{++}$ and $\gamma \in \mathbb{R}$ such that*

$$V_i(t_i) = \beta V_i(t_i') + \gamma e,$$

where e is the unit vector in Δ^{K-1}.

Pseudo-type diversity is a generalization of the type diversity condition for a

[8] This is done for simplicity. If A were an arbitrary separable space, we would work with its countable dense subset. The reader is referred to Section 6 of Abreu and Sen (1991) or to Duggan (1997) for more details.

standard Bayesian environment, used in Serrano and Vohra (2005). The reader is referred to that paper to find an appraisal of the connections of type diversity with the conditions of interim value distinguished types (Palfrey and Srivastava (1993, definition 6.3)), incentive consistency (Duggan (1997)), and with the algorithm behind measurability due to Abreu and Matsushima (1992). We will have more to say about the latter connection in the next sections. [9]

Remark: Pseudo-TD is generically satisfied in the space of pseudo-interim preferences over pure outcomes if $|A| \geq 3$. As noted above, when we consider a consistent type space $\mathcal{T}$ in which $\mathcal{T}_i = \Theta_i$ and Q_i is a singleton for each agent $i \in N$, pseudo-TD is reduced to TD of Serrano and Vohra (2005).

For every consistent type space $\mathcal{T}$, define $U_i^k(\tau_i)$ to be the interim expected utility of agent i of type τ_i for the constant SCF which assigns a_k in each state in T, i.e.,

$$U_i^k(\tau_i) = \sum_{\tau_{-i} \in \mathcal{T}_{-i}} \pi_i(\tau_{-i}|\tau_i) u_i(a_k; \hat{\theta}(\tau_i, \tau_{-i})).$$

So, the condition of TD would ask that no two types of an agent can be found for whom these vectors are positive affine transformations of one another. The next lemma explains how to go from pseudo-TD to TD:

Lemma 2 *Suppose that an environment $\mathcal{E}$ satisfies pseudo-TD. Then, for any consistent type space $\mathcal{T}$, there do not exist $i \in N$, $\tau_i, \tau_i' \in \mathcal{T}_i$ with $\hat{t}_i(\tau_i) \neq \hat{t}_i(\tau_i')$, $\beta > 0$, and $\gamma \in \mathbb{R}$ such that*

$$U_i(\tau_i) = \beta U_i(\tau_i') + \gamma e$$

where e is the unit vector in Δ^{K-1}.

Proof of Lemma 2: Fix an arbitrary consistent type space $\mathcal{T}$. As it will become clear, the argument does not depend on any particular type space consistent with the original environment $\mathcal{E}$. Consider agent i of type τ_i. Let $\hat{t}_i(\tau_i) \equiv t_i$ and $\hat{\theta}_i(\tau_i) \equiv \theta_i$.

[9] If A is a separable metric space, let $A^* = \{a_1, a_2, \ldots\}$ be a countable dense subset of A. Now, we can define

$$V_i(t_i) = (V_i^k(t_i))_{k=1}^{\infty} \in \mathbb{R}^{\infty}$$

We also define e as the countable unit base in A with $\|e\| = 1$. With these qualifications, pseudo-TD is also well defined for separable metric spaces.

We claim that $U_i^k(\tau_i) = V_i^k(t_i)$ for each $k = 1, \ldots, K$.

$$\begin{aligned}
U_i^k(\tau_i) &= \sum_{\tau_{-i} \in \mathcal{T}_{-i}} \pi_i(\tau_{-i}|\tau_i) u_i(a_k; \hat{\theta}(\tau_i, \tau_{-i})) \\
&= \sum_{t_{-i} \in T_{-i}} \sum_{\tau_{-i}:\ \hat{t}_{-i}(\tau_{-i}) = t_{-i}} \pi_i(\tau_{-i}|\tau_i) u_i(a_k; \hat{\theta}(\tau_i, \tau_{-i})) \\
&= \sum_{t_{-i} \in T_{-i}} u_i(a_k; \hat{\theta}(t_i, t_{-i})) \sum_{\tau_{-i}:\ \hat{t}_{-i}(\tau_{-i}) = t_{-i}} \pi_i(\tau_{-i}|\tau_i) \ \ (\because \ \hat{t}_i(\tau_i) = t_i) \\
&= \sum_{t_{-i} \in T_{-i}} q_i(\hat{\theta}_{-i}(t_{-i})|\hat{\theta}_i(t_i)) u_i(a_k; \hat{\theta}(t_i, t_{-i})) \ \ (\because \ t_i = (\hat{\theta}_i(t_i), q_i)) \\
&= \sum_{\theta_{-i} \in \Theta_{-i}} q_i(\theta_{-i}|\theta_i) u_i(a_k; \theta_i, \theta_{-i}) \ \ (\because \ \hat{\theta}_i(t_i) = \theta_i) \\
&= V_i^k(t_i) \ \ (\because \ t_i \equiv (\theta_i, q_i)).
\end{aligned}$$

Thus, we obtain $U_i^k(\tau_i) = V_i^k(t_i)$ whenever $\hat{t}_i(\tau_i) = t_i$. Similarly, consider agent i of type τ_i'. Let $\hat{t}_i(\tau_i') \equiv t_i'$ and $\hat{\theta}_i(\tau_i') \equiv \theta_i'$. Then, by the same argument, we obtain $U_i^k(\tau_i') = V_i^k(t_i')$ whenever $\hat{t}_i(\tau_i') = t_i'$. Having established this, pseudo-TD takes care of the rest of the argument. ■

In environments satisfying pseudo-NTI and pseudo-TD, we next show the following critical lemma, a generalization of Lemma 1 in Serrano and Vohra (2005).

Lemma 3 *Suppose an environment $\mathcal{E}$ satisfies pseudo-NTI and pseudo-TD. Then there exist constant SCFs $\left((\ell_i(t_i))_{t_i \in T_i}\right)_{i \in N}$ such that for every $i \in N$ and $t_i, t_i' \in T_i$ with $t_i \neq t_i'$,*

$$V_i(\ell_i(t_i)|t_i) > V_i(\ell_i(t_i')|t_i).$$

Remark: All is needed for this lemma is the assumption that the individual preferences over lotteries are *monotone* in the sense that any shift of probability weight from a less preferred to a more preferred pure alternative yields a lottery which is preferred. The axiom that preferences are monotone is, of course, much weaker than the independence axiom, and is implied by the von Neumann-Morgenstern utility representation.

Proof of Lemma 3: Consider the constant SCF $\bar{x}$, which prescribes in each state the lottery $\bar{x}$, assigning equal probability to each alternative in A, i.e., $\bar{x}(t) = (1/K, \ldots, 1/K)$ for all $t \in T$. We will use induction on the number of pseudo-types of agent i.

First, we show that for $i \in N$, and for two pseudo-types $t_i, t_i^{'} \in T_i$ with $t_i \neq t_i^{'}$, there exist constant SCFs x and $x^{'}$, close to $\bar{x}$, such that

$$V_i(x|t_i) > V_i(x^{'}|t_i) \quad \text{and} \quad V_i(x^{'}|t_i^{'}) > V_i(x|t_i^{'}). \tag{1}$$

The interim indifference curve of agent i of pseudo-type t_i through $\bar{x}$ is described by a hyperplane, H, in $\mathbb{R}_+^{K-1}$:

$$H = \left\{ (x_1, \ldots, x_{K-1}) \in \mathbb{R}_+^{K-1} \middle| \sum_{k=1}^{K-1} p_k(t_i) x_k = \bar{u} \right\},$$

where $p_k(t_i) = (V_i^k(t_i) - V_i^K(t_i))$ for $k = 1, \ldots, K-1$.

Let $p(t_i) = (p_1(t_i), \ldots, p_{K-1}(t_i)) \in \mathbb{R}^{K-1}$. Consider the interim indifference hyperplane through $\bar{x}$ of agent i of pseudo-type $t_i^{'}$ where $t_i \neq t_i^{'}$:

$$H^{'} = \left\{ (x_1, \ldots, x_{K-1}) \in \mathbb{R}_+^{K-1} \middle| \sum_{k=1}^{K-1} p_k(t_i^{'}) x_k = \bar{u}^{'} \right\},$$

Given pseudo-NTI, we must have $p(t_i) \neq \nvDash$ and $p(t_i^{'}) \neq \nvDash$. We claim that $p(t_i) \neq cp(t_i^{'})$ for any $c > 0$. Suppose not; that is, there is $c > 0$ such that $p(t_i) = cp(t_i^{'})$. This implies that $V_i(t_i) = cV_i(t_i^{'}) + \gamma e$, which contradicts pseudo-TD. Thus, either $p(t_i) = cp(t_i^{'})$ where $c < 0$ or there does not exist $c \neq 0$ such that $p(t_i) = cp(t_i^{'})$. In the former case, it is easy to see (using pseudo-NTI) that any point which lies above H must be below $H^{'}$ and, choosing two points (one above H and one below it) close to $\bar{x}$, one finds constant SCFs which satisfy (1). In the latter case, it is clear that we can choose two constant SCFs which satisfy (1).

Now, according to the induction hypothesis, suppose that for the first $|T_i| - 1$ pseudo-types of agent i, i.e., for all $t_i \in T_i \setminus \{t_i^0\}$, we have been able to find $|T_i| - 1$ constant SCFs near $\bar{x}$, say $x(t_i)$, such that for every $t_i \in T_i \setminus \{t_i^0\}$, $V_i(x(t_i)|t_i) >$

$V_i(x(t_i^{'})|t_i)$ for every $t_i^{'} \in T_i \setminus \{t_i^0, t_i\}$. Consider pseudo-type t_i^0. Choose the constant SCF among the collection $(x(t_i))_{t_i \in T_i \setminus \{t_i^0\}}$ that is ranked highest by pseudo-type t_i^0 (without loss of generality, there is only one). Call it $x(t_i)$. By arguments similar to the ones in the previous paragraph, because of pseudo-NTI and pseudo-TD, one can find a constant SCF near $x(t_i)$, call it $x(t_i^0)$, such that pseudo-types t_i and t_i^0 satisfy (1). Finally, since all inequalities concerning the other pseudo-types and their associated SCFs are strict, $x(t_i^0)$ can be chosen so that the collection of constant SCFs $(x(t_i))_{t_i \in T_i}$ satisfy all the inequalities in the statement of the lemma, so the proof is complete.[10] ■

Corollary 1 *Suppose an environment $\mathcal{E}$ satisfies pseudo-NTI and pseudo-TD. Then there exist constant SCFs $\left((\ell_i(t_i))_{t_i \in T_i}\right)_{i \in N}$ such that for any consistent type space $\mathcal{T}$ in which for every $i \in N$ and $\tau_i, \tau_i^{'} \in \mathcal{T}_i$ with $\hat{t}_i(\tau_i) \neq \hat{t}_i(\tau_i^{'})$,*

$$U_i(\ell_i(\hat{t}_i(\tau_i))|\tau_i) > U_i(\ell_i(\hat{t}_i(\tau_i^{'}))|\tau_i).$$

Proof of Corollary 1: This follows directly from Lemmas 2 and 3. ■

3 A Robust Canonical Mechanism

This section introduces a mechanism that will be used to obtain a very permissive robust virtual implementation result over all consistent type spaces. Among its virtues, one should stress its finiteness, so that best replies are always well defined.

The mechanism $\Gamma = (M, g)$ uses the collection of constant SCFs ℓ_i of Lemma 3.[11] The construction is as follows: Every player i makes $(J+1)$ simultaneous announcements, each of which is of his own pseudo-type

$$M_i = M_i^0 \times M_i^1 \times \cdots \times M_i^J = \underbrace{T_i \times T_i \times \cdots \times T_i}_{J+1}.$$

[10] If A is a separable metric space, the modification we must make to the previous argument is the way we define the lottery $\bar{x}(t)$:

$$\bar{x}(t) = (\bar{x}_k(t))_{k=1}^{\infty}$$

where $\bar{x}_k(t) = (1-\delta)\delta^{k-1}$, and $0 < \delta < 1$.

[11] It is inspired by the heuristic section of Abreu and Matsushima (1992) that precedes their formal analysis. We dispense with two important assumptions made there: private values and the existence of a nummeraire.

Denote

$$\begin{aligned} m_i &= \left(m_i^0, \ldots, m_i^J\right) \in M_i,\ m_i^s \in M_i^s\ \forall s = 0, \ldots, J \\ m &= \left(m^0, \ldots, m^J\right) \in M,\ m^s = (m_i^s)_{i \in N} \in M^s = \times_{i \in N} M_i^s \end{aligned}$$

We introduce the following *bribe/punishment* lottery to reward a coherent announcement from each agent:

$$\xi(i,m) = \begin{cases} \arg\min_{\ell_i(t_i) \in \{\ell_i(t_i)\}_{t_i \in T_i}} \{V_i(\ell_i(t_i)|m_i^0)\} & \text{if } \exists j \in \{1, \ldots, J\} \text{ s.t.} \\ & m_i^j \neq m_i^0 \text{ for } i \in N \text{ and} \\ & m^s = m^0\ \forall\ s \in \{0, \ldots, j-1\}. \\ \\ \ell_i(m_i^0) & \text{otherwise} \end{cases}$$

We call it bribe/punishment lottery because the agent gets his best ℓ_i, given m_i^0, unless the agent changes his announcement before any change in announcements is observed in the previous rounds, in which case he gets the worst ℓ_i, given m_i^0.

Define an SCF

$$\ell(t) = \frac{1}{n} \sum_{i \in N} \ell_i(t_i).$$

Given an SCF f, for any profile of agents' messages m, the *outcome function* of the mechanism is

$$g(m) = \varepsilon \ell(m^0) + \frac{\varepsilon^2}{n} \sum_{i \in N} \xi(i,m) + \frac{1 - \varepsilon - \varepsilon^2}{J} \sum_{s=1}^{J} \left\{\varepsilon^2 \ell(m^s) + \left(1 - \varepsilon^2\right) f(m^s)\right\}$$

where ε is small and strictly positive.

This outcome function has three terms: the first, weighted by a probability of ε, depends only on m^0 and consists of the SCFs from Section 2 that induce the separation of types; the second, weighted by ε^2, is the bribe/punishment lottery we have just constructed; the third term, having the remaining weight, depends on the rest of the announcements $m^1, \ldots, m^J$ and consists of the (slightly modified) SCF being implemented.

Equivalently,

$$g(m) = \varepsilon\ell(m^0) + \frac{\varepsilon^2}{n}\sum_{i\in N}\xi(i,m) + \frac{1-\varepsilon-\varepsilon^2}{J}\sum_{s=1}^{J}\tilde{f}(m^s),$$

where

$$\tilde{f}(m^s) = \varepsilon^2\ell(m^s) + (1-\varepsilon^2)f(m^s).$$

Note that if f satisfies incentive compatibility, $\tilde{f}$ satisfies strict incentive compatibility. This is because of the addition of the ℓ_i terms. Besides, $\tilde{f}$ is close to f for small $\varepsilon > 0$.

4 The Main Result

Fix an arbitrary consistent type space $\mathcal{T}$. Let σ be an iteratively undominated strategy profile. Recall that $\sigma_i : \mathcal{T}_i \to M_i$. Denote strategies for players by

$$\begin{aligned}\sigma_i &= \left(\sigma_i^0, \sigma_i^1, \ldots, \sigma_i^J\right),\ \sigma_i^s : \mathcal{T}_i \to M_i^s,\\ \sigma &= \left(\sigma^0, \sigma^1, \ldots, \sigma^J\right),\ \sigma^s : \mathcal{T} \to M^s.\end{aligned}$$

Theorem 1 *Suppose an environment $\mathcal{E}$ satisfies pseudo-NTI and pseudo-TD. If an SCF f is incentive compatible for every consistent type space $\mathcal{T}$, it is robustly virtually implementable in iteratively undominated strategies.*

Proof of Theorem 1: Fix an arbitrary consistent type space $\mathcal{T}$. It will be clear that the argument does not depend on $\mathcal{T}$, as long as it is consistent with the pseudo-type space. The proof consists of two claims using the mechanism of the previous section.

Claim 1.1: Suppose that σ is an iteratively undominated strategy profile of the mechanism Γ. Then, $\sigma_i^0(\tau_i) = \hat{t}_i(\tau_i)$ for all $i \in N$ and $\tau_i \in \mathcal{T}_i$.

Proof of Claim 1.1: We can choose $\varepsilon > 0$ small enough so that

$$\min_{i\in N, t_i\in T_i, t_i'\neq t_i}\left\{V_i(\ell_i(t_i)|t_i) - V_i(\ell_i(t_i')|t_i)\right\} > \varepsilon \max_{i\in N, t_i\in T_i, t_i'\neq t_i}\left\{V_i(\ell_i(t_i)|t_i) - V_i(\ell_i(t_i')|t_i)\right\}$$

The above inequality is well defined when N is finite and T_i is finite for every $i \in N$. Define $\eta_0, \eta_1 > 0$ with $\eta_0 < \eta_1$ as follows:

$$\begin{aligned}
\eta_0 &\equiv \frac{\varepsilon^2}{n} \min_{i \in N, t_i \in T_i, t_i' \neq t_i} \left\{ V_i(\ell_i(t_i)|t_i) - V_i(\ell_i(t_i')|t_i) \right\} > 0 \\
\eta_1 &\equiv \frac{\varepsilon^2}{n} \max_{i \in N, t_i \in T_i, t_i' \neq t_i} \left\{ V_i(\ell_i(t_i)|t_i) - V_i(\ell_i(t_i')|t_i) \right\} > 0
\end{aligned}$$

Here, η_0 and η_1 are the minimal and maximal effects on pseudo-interim expected utility associated with the "bribe/punishment" lottery, respectively. Then, by our choice of ε and the definition of η_1, for any $i \in N$, we have

$$\frac{\varepsilon}{n} \left\{ V_i(\ell_i(t_i)|t_i) - V_i(\ell_i(t_i')|t_i) \right\} > \eta_1 \ \forall \ t_i \in T_i, t_i' \in T_i \backslash \{t_i\}.$$

Note that $\varepsilon, \eta_0, \eta_1$ are chosen independently of the choice of any particular consistent type space.

Recall the outcome function of the mechanism, and notice that announcement m_i^0 affects only the first term and possibly the second through the "bribe/punishment" lottery. According to the last inequality, the payoff loss from misreporting one's pseudo-type in m_i^0 exceeds the maximum possible gain from the second term, whatever strategies are used by the other agents. Thus, player i will be strictly better off by telling the truth in the 0th announcement, even if he were to misrepresent the rest of his announcements.

Formally, we argue by contradiction. Let σ be a strategy profile such that $\sigma_i^0(\tau_i) = t_i \neq \hat{t}_i(\tau_i)$ for some player i of some type τ_i.

Define $\hat{\sigma}_i$ as follows:

$$\begin{aligned}
\hat{\sigma}_i^s &= \sigma_i^s \ \forall s \geq 1, \\
\hat{\sigma}_i^0(\tau_i') &= \sigma_i^0(\tau_i') \ \forall \tau_i' \neq \tau_i, \\
\text{and } \hat{\sigma}_i^0(\tau_i) &= \hat{t}_i(\tau_i).
\end{aligned}$$

We compare below the interim utilities of agent i of type τ_i when he employs σ_i and $\hat{\sigma}_i$ against any $\tilde{\sigma}_{-i} \in \Sigma_{-i}$:

$$\begin{aligned} & U_i(g \circ (\hat{\sigma}_i, \tilde{\sigma}_{-i})|\tau_i) \\ = & \frac{\varepsilon}{n} V_i(\ell_i(\hat{t}_i(\tau_i))|\hat{t}_i(\tau_i)) + \frac{\varepsilon^2}{n} V_i(\xi(i, \hat{\sigma}_i, \tilde{\sigma}_{-i})|\hat{t}_i(\tau_i)) + \lambda \\ > & \frac{\varepsilon}{n} V_i(\ell_i(t_i)|\hat{t}_i(\tau_i)) + \frac{\varepsilon^2}{n} V_i(\xi(i, \sigma_i, \tilde{\sigma}_{-i})|\hat{t}_i(\tau_i)) + \lambda \\ = & U_i(g \circ (\sigma_i, \tilde{\sigma}_{-i})|\tau_i), \end{aligned}$$

where λ is a shorthand that denotes the rest of terms, which are the same in both expressions. Thus, $\hat{\sigma}_i$ strictly dominates σ_i. ■

Claim 1.2: For every $i \in N$, let σ_i be an iteratively undominated strategy. Suppose that $\sigma_i^s(\tau_i) = \hat{t}_i(\tau_i)$ for all $\tau_i \in \mathcal{T}_i$ and $s \in \{0, \ldots, j\}$, where $0 \leq j \leq J-1$. Then

$$\sigma_i^{j+1}(\tau_i) = \hat{t}_i(\tau_i) \text{ for all } i \in N \text{ and all } \tau_i \in \mathcal{T}_i.$$

Proof of Claim 1.2: We need some additional pieces of notation for the proof. A deception is a profile of functions, $\alpha = (\alpha_i)_{i \in N}$, where $\alpha_i : T_i \to T_i$. Consider the SCF $\tilde{f}$, a pseudo-type $t_i \in T_i$, and a deception α with the property that $\alpha_i(t_i) \neq t_i$ and $\alpha_{i'}(t_{i'}) \neq t_{i'}$ for some player $i' \neq i$ of type $t_{i'}$. Let $\tilde{f} \circ \alpha(t) = \tilde{f}(\alpha(t))$ for all $t \in T$. Define the following:

$$\begin{aligned} \gamma_i(t_i) &\equiv \max_{\alpha} V_i\left(\tilde{f} \circ \alpha \middle| t_i\right) - \min_{\alpha} V_i\left(\tilde{f} \circ \alpha \middle| t_i\right) \\ \gamma_i &\equiv \max_{t_i \in T_i} \gamma_i(t_i) \\ \gamma &\equiv \max_{i \in N} \gamma_i \end{aligned}$$

The number γ is well defined because T_i is finite for every $i \in N$.[12]

Suppose, by way of contradiction, that $\sigma_i^{j+1}(\tau_i) \neq \hat{t}_i(\tau_i)$ for some player i of some type τ_i.

[12]There is an implicit assumption here that the maximum and minimum terms in $\gamma_i(t_i)$ are not the same for every i and every t_i, which will be true for almost all values of ε – see the definition of $\tilde{f}$.

Define $\bar{\sigma}_i$ such that

$$\begin{aligned} \bar{\sigma}_i^s &= \sigma_i^s \ \forall s \neq j+1, \\ \bar{\sigma}_i^{j+1}(\tau_i^{'}) &= \sigma_i^{j+1}(\tau_i^{'}) \ \forall \tau_i^{'} \neq \tau_i, \\ \text{and } \bar{\sigma}_i^{j+1}(\tau_i) &= \hat{t}_i(\tau_i). \end{aligned}$$

Under the induction hypothesis, if $\sigma_{i'}^{j+1}(\tau_{i'}) = \hat{t}_{i'}(\tau_{i'})$ for all $i' \neq i$ and all $\tau_{i'} \in \mathcal{T}_{i'}$, then, by strict incentive compatibility of $\tilde{f}$, $\bar{\sigma}_i$ yields higher payoff than σ_i in the $j+1$-st term of the third part of the outcome function. In addition, the "bribe/punishment" second term cannot get worse by using $\bar{\sigma}_i$ instead of σ_i. Thus, in this case, $\bar{\sigma}_i$ has a higher expected payoff than σ_i.

On the other hand, suppose that $\sigma_{i'}^{j+1}(\tau_{i'}) \neq \hat{t}_{i'}(\tau_{i'})$ for some player $i' \neq i$ of type $\tau_{i'} \in \mathcal{T}_{i'}$. Then, by construction of γ, for any σ_{-i} under the inductive hypothesis, we have

$$\gamma \geq U_i(\tilde{f} \circ \sigma^{j+1} | \tau_i) - U_i(\tilde{f} \circ (\bar{\sigma}_i^{j+1}, \sigma_{-i}^{j+1}) | \tau_i).$$

We choose J large enough so that

$$\eta_0 > \frac{1-\varepsilon}{J}\gamma \geq \frac{1-\varepsilon}{J}\left\{U_i(\tilde{f} \circ \sigma^{j+1} | \tau_i) - U_i(\tilde{f} \circ (\bar{\sigma}_i^{j+1}, \sigma_{-i}^{j+1}) | \tau_i)\right\}.$$

Then, by improving his payoff in the "bribe/punishment" term, $\bar{\sigma}_i$ yields higher payoff than σ_i. That is, for any σ_{-i} under the inductive hypothesis, we have

$$U_i(g \circ (\bar{\sigma}_i, \sigma_{-i}) | \tau_i) > U_i(g \circ \sigma | \tau_i)$$

In other words, under the inductive hypothesis, it is always better for player i of type τ_i to wait for one more round to misrepresent his type so that other players misrepresent their type first, thereby avoiding the punishment involved in the second term of the outcome function. This, however, contradicts our hypothesis that σ_i is an iteratively undominated strategy. ■

Claims 1.1 and 1.2 together show that there is a unique iteratively undominated strategy profile σ with the property that $\sigma_i^s(\tau_i) = \hat{t}_i(\tau_i) = t_i$ for every $i \in N,\ \tau_i \in \mathcal{T}_i$,

and $s \in \{0, 1, \ldots, J\}$. The resulting outcome is

$$(1-\varepsilon^2)\left(1-\varepsilon-\varepsilon^2\right) f(t) + \frac{\varepsilon+\varepsilon^2+(1-\varepsilon-\varepsilon^2)\varepsilon^2}{n} \sum_{i \in N} \ell_i(t_i).$$

This outcome is arbitrarily close to $f(t)$ for every $t \in T$ when $\varepsilon > 0$ is chosen to be small enough. This completes the proof of Theorem 1. ■

5 A-M Measurability as a Necessary Condition

What we have shown so far is that *generically* robust virtual implementation in iteratively undominated strategies is as successful as it can possibly be. That is, in environments satisfying pseudo-type diversity, an SCF that is incentive compatible on every consistent type space is robustly virtually implementable in iteratively undominated strategies. In an important paper, Abreu and Matsushima (1992) uncovered a condition that they termed *measurability* (we shall refer to it from now on as A-M measurability) that was necessary for virtual implementation in iteratively undominated strategies over a standard environment that fixes a Bayesian type space. In this section we revisit the A-M measurability condition by applying it to our robust implementation analysis. In the process, a connection with pseudo-type diversity will also be explained.

Denote by Ψ_i a *partition* of the set of pseudo-types T_i, where ψ_i is a generic element of Ψ_i and $\Pi_i(t_i)$ is the element of Ψ_i that includes pseudo-type t_i. Let $\Psi = \times_{i \in N} \Psi_i$ and $\psi = \times_{i \in N} \psi_i$.

Definition 10 *An SCF f is **measurable with respect to** Ψ if, for every $i \in N$ and every $t_i, t_i' \in T_i$, whenever $\Pi_i(t_i) = \Pi_i(t_i')$,*

$$f(t_i, t_{-i}) = f(t_i', t_{-i}) \ \ \forall t_{-i} \in T_{-i}.$$

Measurability of f with respect to Ψ implies that for any player i, f does not distinguish between any pair of pseudo-types in the same cell of the partition Ψ_i.

Definition 11 *Let $\mathcal{T}$ be a consistent type space. A strategy σ_i for player i is **mea-***

surable with respect to Ψ_i *if for every* $\tau_i, \tau_i^{'} \in \mathcal{T}_i$,

$$\Pi_i(\hat{t}_i(\tau_i)) = \Pi_i(\hat{t}_i(\tau_i^{'})) \Longrightarrow \sigma_i(\tau_i) = \sigma_i(\tau_i^{'}).$$

A strategy profile σ *is* ***measurable with respect to*** Ψ *if, for every* $i \in N$, σ_i *is measurable with respect to* Ψ_i.

For every $i \in N$, $t_i, t_i^{'} \in T_i$, and $(n-1)$ tuple of partitions Ψ_{-i}, we say that t_i is *equivalent* to $t_i^{'}$ with respect to Ψ_{-i} if, for every f and every $\tilde{f}$ which are measurable with respect to $T_i \times \Psi_{-i}$,

$$V_i(f|t_i) \geq V_i(\tilde{f}|t_i) \Longleftrightarrow V_i(f|t_i^{'}) \geq V_i(\tilde{f}|t_i^{'}).$$

Fix a consistent type space $\mathcal{T}$. Then, we say that τ_i is equivalent to $\tau_i^{'}$ with respect to Ψ_{-i} if, for every f and $\tilde{f}$ that are measurable with respect to $T_i \times \Psi_{-i}$,

$$U_i(f|\tau_i) \geq U_i(\tilde{f}|\tau_i) \Longleftrightarrow U_i(f|\tau_i^{'}) \geq U_i(\tilde{f}|\tau_i^{'}).$$

Lemma 4 *Let* $\mathcal{T}$ *be any type space consistent with the original environment. Then, type* τ_i *is equivalent to type* $\tau_i^{'}$ *with respect to* Ψ_{-i} *whenever* $\hat{t}_i(\tau_i)$ *is equivalent to* $\hat{t}_i(\tau_i^{'})$ *with respect to* Ψ_{-i}.

Proof of Lemma 4: Fix an arbitrary consistent type space $\mathcal{T}$. Let $t_i \equiv \hat{t}_i(\tau_i)$ and $t_i^{'} \equiv \hat{t}_i(\tau_i^{'})$. Consider an arbitrary type τ_i and an arbitrary SCF $f : T \rightarrow \Delta(A)$. By arguments identical to those used in the proof of Lemma 1, one can show that $U_i(f|\tau_i) = V_i(f|t_i)$.

Consider arbitrary SCFs f and $\tilde{f}$ that are measurable with respect to $T_i \times \Psi_{-i}$. Then, the hypothesis that t_i is equivalent to $t_i^{'}$ with respect to Ψ_{-i} implies

$$V_i(f|t_i) \geq V_i(\tilde{f}|t_i) \Longleftrightarrow V_i(f|t_i^{'}) \geq V_i(\tilde{f}|t_i^{'}).$$

With the obtained equivalence that $U_i(f|\tau_i) = V_i(f|t_i)$ and $U_i(f|\tau_i^{'}) = V_i(f|t_i^{'})$ for any SCF f, we can conclude

$$U_i(f|\tau_i) \geq U_i(\tilde{f}|\tau_i) \Longleftrightarrow U_i(f|\tau_i^{'}) \geq U_i(\tilde{f}|\tau_i^{'}).$$

This implies that τ_i is equivalent to $\tau_i^{'}$ with respect to Ψ_{-i}. ■

Fix an arbitrary consistent type space $\mathcal{T}$. Suppose that player i believes that every SCF is measurable with respect to $T_i \times \Psi_{-i}$. Assume further that τ_i is equivalent to $\tau_i^{'}$ with respect to Ψ_{-i}. Then, player i's interim expected utility under type τ_i is exactly the same as under type $\tau_i^{'}$ when evaluating any SCF.

Let $\rho_i(t_i, \Psi_{-i})$ be the set of all elements of T_i that are equivalent to t_i with respect to Ψ_{-i}, and let

$$R_i(\Psi_{-i}) = \{\rho_i(t_i, \Psi_{-i}) \subset T_i |\ t_i \in T_i\}.$$

Note that $R_i(\Psi_{-i})$ forms an equivalence class on T_i, that is, constitutes a partition of T_i. We define an infinite sequence of n-tuples of partitions, $\{\Psi^h\}_{h=0}^{\infty}$, where $\Psi^h = \times_{i \in N} \Psi_i^h$ in the following way. For every $i \in N$,

$$\Psi_i^0 = \{T_i\},$$

and recursively, for every $i \in N$ and every $h \geq 1$,

$$\Psi_i^h = R_i(\Psi_{-i}^{h-1}).$$

Note that for every $h \geq 0$, Ψ_i^{h+1} is the same as, or finer than, Ψ_i^h. Define Ψ^* as follows:

$$\Psi^* \equiv \bigcup_{h=0}^{\infty} \Psi^h.$$

Since T_i is finite for each agent $i \in N$, Lemma 4 guarantees that there exists a positive integer L such that $\Psi^h = \Psi^L$ for any $h \geq L$. We denote $\Psi^* = \Psi^L$.

Definition 12 *An SCF f is **A-M measurable** if it is measurable with respect to Ψ^*.*

Note how the partitions Ψ^0, Ψ^1, ..., and hence, the final partition Ψ^* used in A-M measurability are really nothing but a property of the environment. The aim is to "treat equally" those pseudo-types that are "indistinguishable" according to their interim preferences. Thus, we start considering constant SCFs, i.e., SCFs that are

measurable with respect to the coarsest possible partition, and we separate pseudo-types who have different interim preferences over this class of SCFs. This gives us a new partition of the set of pseudo-types for each agent (iteration 1). Next, we consider SCFs measurable with respect to these new partitions, and ask the same question: are there pseudo-types that, having the same preferences over constant SCFs, now can be separated because they exhibit different interim preferences over the enlarged class of SCFs considered? If the answer is No, the process ends and we have found Ψ^*. If it is Yes, we proceed to make the induced finer partition of each set of pseudo-types (iteration 2), and so on. The process ends after a finite number of steps with the identification of Ψ^*, which provides the maximum possible degree of pseudo-type separation or distinguishability in terms of interim preferences. A-M measurability simply asks that the SCF not distinguish between different pseudo-types that are "indistinguishable" according to Ψ^*.

When a consistent type space $\mathcal{T}$ satisfies the properties that $\mathcal{T}_i = \Theta_i$ and Q_i is a singleton for each $i \in N$, A-M measurability is reduced to the measurability proposed by Abreu and Matsushima (1992).

Define

$$F = \{h \mid h(t) \text{ is a degenerate lottery for all } t \in T\}.$$

Recall that T_i is finite for every $i \in N$. Assume also that A is finite.[13] Then, F becomes a finite functional space. Define also

$$F(\Psi) = \{h \in F| \ h \text{ is measurable with respect to } \Psi\}.$$

Let $|F(T_i \times \Psi_{-i})| = K$.[14] Define $V_i^k(t_i, \Psi_{-i})$ to be the interim expected utility of agent i of pseudo-type t_i for each SCF $f^k \in F(T_i \times \Psi_{-i})$, i.e.,

$$V_i^k(t_i, \Psi_{-i}) = \sum_{\theta_{-i} \in \Theta_{-i}} \sum_{q_{-i} \in Q_{-i}} q_i(\theta_{-i}|\hat{\theta}_i(t_i)) u_i(f^k(t_i, \theta_{-i}, q_{-i}); \hat{\theta}_i(t_i), \theta_{-i})).$$

Let $V_i(t_i, \Psi_{-i}) = (V_i^1(t_i, \Psi_{-i}), \ldots, V_i^K(t_i, \Psi_{-i}))$.

[13] If A were a separable space, we would work with its countable dense subset.

[14] This is a slight abuse of notation, since K was defined in previous sections as the finite number of alternatives in the set A. In part, we choose to use the same symbol here to enhance the parallels across the arguments in the different sections. Also, it should not cause any confusion.

The next lemma follows simply from the definitions of $F(\Psi)$ and of equivalent types. Its proof is omitted:

Lemma 5 *Assume that A is finite. Then, t_i is equivalent to $t_i^{'}$ with respect to Ψ_{-i} if and only if there exist $\beta > 0$ and $\gamma \in \mathbb{R}$ such that*

$$V_i(t_i, \Psi_{-i}) = \beta V_i(t_i^{'}, \Psi_{-i}) + \gamma e,$$

where e is the unit vector in Δ^{K-1}.

The following is a characterization of pseudo-TD in terms of the measurability construction:

Corollary 2 *An environment $\mathcal{E}$ satisfies pseudo-NTI and pseudo-TD if and only if there do not exist $i \in N$ and $\tau_i, \tau_i^{'} \in \mathcal{T}_i$ with $t_i = \hat{t}_i(\tau_i) \neq \hat{t}_i(\tau_i^{'}) = t_i^{'}$ such that t_i is equivalent to $t_i^{'}$ with respect to Ψ_{-i}^0 for every consistent type space $\mathcal{T}$. It follows that $\Psi_i^1 = T_i$ for each agent $i \in N$, and $\Psi^* = T$ in every consistent type space $\mathcal{T}$.*

In light of Corollary 2, one can make the following useful observation (see Serrano and Vohra (2005) for a similar assertion concerning TD):

Lemma 6 (TD and NTI $\Rightarrow$ A-M measurability) *Suppose an environment satisfies pseudo-NTI and pseudo-TD. Then, **every** SCF is A-M measurable.*

That is, if the environment satisfies pseudo-NTI and pseudo-TD, the algorithm that separates types in the definition of measurability arrives at the finest partition at the first round. As already said, Abreu and Matsushima (1992) show that A-M measurability is a necessary condition for virtual implementation in iteratively undominated strategies. We adapt their proof to our setup:

Proposition 2 *If an SCF f is robustly virtually implementable in iteratively undominated strategies, then it is A-M measurable.*

Proof of Proposition 2: Since f is robustly virtually implementable in iteratively undominated strategies, there exists f^{ε} that is exactly implementable in iteratively undominated strategies and $d(f, f^{\varepsilon}) < \varepsilon$ for $\varepsilon > 0$ small for any consistent type space $\mathcal{T}$. Consider a mechanism $\Gamma = (M, g)$ which exactly implements the SCF

f^ε in iteratively undominated strategies for any consistent type space $\mathcal{T}$. Fix an arbitrary consistent type space $\mathcal{T}$ and for each $h \geq 1$, let $\mathcal{K}^h = \times_{i \in N} \mathcal{K}_i^h$ be the sets of iteratively undominated strategies at the h-th round of iterative removal for the type space $\mathcal{T}$.

Consider an arbitrary "constant" strategy profile $\sigma[0] \in \mathcal{K}^0$ which is measurable with respect to $\times_{i \in N}\{T_i\}$. Then, either $g(\sigma[0]) = f^\varepsilon$, which is then constant, i.e., measurable with respect to $\times_{i \in N}\{T_i\}$, and hence we are done because it is A-M measurable a fortiori (i.e., measurable with respect to Ψ^*), or $g(\sigma[0]) \neq f^\varepsilon$.

In this case, by the definition of Ψ^1 and our hypothesis that f^ε is exactly implementable in iteratively undominated strategies for the type space $\mathcal{T}$, it follows that for every $i \in N$, there exists $\sigma_i[1] \in \Sigma_i$ that is a best response to $\sigma_{-i}[0]$ and is measurable with respect to Ψ_i^1. Hence, $\sigma_i[1]$ is not strictly dominated for player i with respect to $\mathcal{K}^0$, that is, $\sigma_i[1] \in \mathcal{K}_i^1$. Again, either $g(\sigma[1]) = f^\varepsilon$, but then f^ε is measurable with respect to Ψ^1, and hence A-M measurable; or $g(\sigma[1]) \neq f^\varepsilon$, in which case at least one type finds his strategy $\sigma_i[1]$ as strictly dominated given $\mathcal{K}^1$, and so on.

Take an arbitrary $h = 2, 3, \ldots$, and suppose that there exists a strategy profile $\sigma[h-1] \in \mathcal{K}^{h-1}$ that is measurable with respect to Ψ^{h-1}. Again, either $g(\sigma[h-1]) = f^\varepsilon$ and we are done, or not. If not, since f^ε is exactly implementable in iteratively undominated strategies for the type space $\mathcal{T}$ by our hypothesis, for every $i \in N$, there exists $\sigma_i[h] \in \Sigma_i$ that is a best response to $\sigma_{-i}[h-1]$ and is measurable with respect to Ψ_i^h. Therefore, $\sigma_i[h]$ is not strictly dominated for player i with respect to $\mathcal{K}^{h-1}$. Hence, for all $h = 0, 1, \ldots$, there exists $\sigma[h] \in \mathcal{K}^h$ that is measurable with respect to Ψ^h.

Let σ^* be an iteratively undominated strategy profile in the implementing game form Γ. Then, the preceding argument implies that σ^* is measurable with respect to Ψ^*. It follows that $f^\varepsilon = g \circ \sigma^*$ is measurable with respect to Ψ^* and therefore, is A-M measurable. Finally, for sufficiently small $\varepsilon > 0$, it follows that f is A-M measurable if and only if f^ε is A-M measurable. Note how the same conclusion obtains regardless of any particular consistent type space $\mathcal{T}$. ■

6 A Characterization of Robust Virtual Implementation

For a fixed type space, Abreu and Matsushima (1992) show that, under an additional assumption essentially similar to quasilinear utilities (Assumption 2 in their paper) and using small fines to punish off-equilibrium behavior, A-M measurability and incentive compatibility are sufficient for virtual implementation in iteratively undominated strategies. In our environments, we also establish that (appropriately reformulated) incentive compatibility and A-M measurability are sufficient as well as necessary for robust virtual implementation. We note that we are not making assumptions equivalent to Abreu and Matsushima's Assumption 2.

Given our results so far – Theorem 1 – we know that A-M measurability is "almost always" a trivial condition, since it can be completely dispensed with in environments satisfying pseudo-TD. For the rest of environments, A-M measurability imposes additional restrictions, and sometimes those restrictions are so severe that only constant SCFs can be virtually implemented (see Serrano and Vohra (2001), Bergemann and Morris (2007)). We turn to formalities now.

Recall the recursive construction behind A-M measurability, and, in particular, the partitions Ψ_i^h for $i \in N$ and $h = 0, 1, \ldots$. For each $i \in N$, $t_i \in T_i$, and $h \geq 0$, let $\Pi_i^h(t_i)$ be the element of Ψ_i^h that includes t_i.

As we will be using a mechanism similar to the one in section 3, our initial task is to construct the first – separating – term of the outcome function. The next lemma provides SCFs that will help us separate pseudo-types, as allowed by the h-th iteration in the measurability construction. It is a generalization of Lemma 3.

Lemma 7 *Suppose an environment $\mathcal{E}$ satisfies pseudo-NTI. Then, for every $i \in N$ and every $h = 1, 2, \ldots, L$, there exist SCFs $x_i^h[\psi_i^h] : T \to \Delta(A)$, which are measurable with respect to $\Psi_i^h \times \Psi_{-i}^{h-1}$, and such that for every $t_i \in T_i$ and $\psi_i^h \in \Psi_i^h \backslash \Pi_i^h(t_i)$,*

$$V_i\left(x_i^h[\Pi_i^h(t_i)]\middle|t_i\right) > V_i\left(x_i^h[\psi_i^h]\middle|t_i\right).$$

Recall that

$$V_i(x_i^h[\cdot]|t_i) \equiv \sum_{\theta_{-i}\in\Theta_{-i}} \sum_{q_{-i}\in Q_{-i}} q_i(\theta_{-i}|\theta_i) u_i(x_i^h[\cdot](\cdot,\theta_{-i},q_{-i});\hat{\theta}_i(t_i),\theta_{-i})).$$

Proof of Lemma 7: Again we write the proof for the case when A is finite.[15] Fix iteration h in the A-M measurability algorithm. Consider the SCF $\bar{x}^h$, which prescribes in each state the lottery $\bar{x}^h$, assigning equal probability to each SCF in $F(\Psi_i^h \times \Psi_{-i}^{h-1})$, the space of degenerate lotteries measurable with respect to $\Psi_i^h \times \Psi_{-i}^{h-1})$. That is,

$$\bar{x}^h(t) = \frac{1}{K^h} f^1(t) + \ldots + \frac{1}{K^h} f^{K^h}(t)$$

for all $t \in T$. Here, $|F(\Psi_i^h \times \Psi_{-i}^{h-1})| = K^h$. By construction, $\bar{x}^h$ is measurable with respect to $\Psi_i^h \times \Psi_{-i}^{h-1}$, and, abusing notation, we can write $\bar{x}^h(t) = \bar{x}^h(\Pi^h(t))$.[16]

We claim that for every $i \in N$, every $t_i, t_i^{'} \in T_i$, with $\Pi_i^h(t_i) \neq \Pi_i^h(t_i^{'})$, there exist SCFs $x_i^h[\Pi_i^h(t_i)]$ and $x_i^h[\Pi_i^h(t_i^{'})]$ that are measurable with respect to $\Psi_i^h \times \Psi_{-i}^{h-1}$, close to $\bar{x}^h$, such that

$$V_i(x_i^h[\Pi_i^h(t_i)]|t_i) > V_i(x_i^h[\Pi_i^h(t_i^{'})]|t_i) \quad \text{and} \quad V_i(x_i^h[\Pi_i^h(t_i^{'})]|t_i^{'}) > V_i(x_i^h[\Pi_i^h(t_i)]|t_i^{'}). \quad (2)$$

We can prove this claim by using the same argument as in Lemma 3. That is, consider the $(K^h - 1)$-dimensional unit simplex, whose extreme points are the elements of the functional space $F(\Psi_i^h \times \Psi_{-i}^{h-1})$. Note how the pseudo-interim expected utility of each extreme point is well defined for each pseudo-type, and thus, one can consider the corresponding hyperplanes as the level curves of such interim utility. By construction of the h-th iteration of measurability, pseudo-types t_i and $t_i^{'}$ can be separated in their interim preferences over SCFs in $F(\Psi_i^h \times \Psi_{-i}^{h-1})$ whenever $\Pi_i^h(t_i) \neq \Pi_i^h(t_i^{'})$. Then, using the argument in the proof of Lemma 3, one can find two SCFs to separate the two pseudo-types as written in (2). The rest of the argument is based on an induction step on the number of elements of Ψ_i^h, exactly as in the proof of Lemma 3. ■

[15] If A were a separable metric space, we would work with its countable dense subset as in footnote 9.

[16] In fact, given the mechanism we construct below, in which agents report atoms of the partition ψ_i^* and not pseudo-types, this will be a convenient way to write the argument of an SCF. Therefore, we shall use this repeatedly in the rest of this section.

The next lemma extends the previous one from pseudo-types to types in a consistent type space:

Lemma 8 *Suppose an environment $\mathcal{E}$ satisfies pseudo-NTI. Then, for every $i \in N$ and every $h = 1, 2, \ldots, L$, there exist SCFs $x_i^h[\psi_i^h] : T \to \Delta(A)$ that are measurable with respect to $\Psi_i^h \times \Psi_{-i}^{h-1}$ such that for every consistent type space $\mathcal{T}$, for every $\tau_i \in \mathcal{T}_i$ with $\hat{t}_i(\tau_i) = t_i \in T_i$ and every $\psi_i^h \in \Psi_i^h \backslash \Pi_i^h(t_i)$,*

$$U_i\left(x_i^h[\Pi_i^h(t_i)]\big|\tau_i\right) > U_i\left(x_i^h[\psi_i^h]\big|\tau_i\right).$$

Recall that

$$U_i(x_i^h[\cdot]|\tau_i) \equiv \sum_{\tau_{-i} \in \mathcal{T}_{-i}} \pi_i(\tau_{-i}|\tau_i) u_i(x_i^h[\cdot](\cdot, \hat{t}_{-i}(\tau_{-i})); \hat{\theta}(\tau_i, \tau_{-i})).$$

Proof of Lemma 8: This follows directly from Lemmas 4 and 7. ■

We are now ready to state and prove the main result of this section:

Theorem 2 (A Characterization of Robust Virtual Implementation) *Suppose an environment $\mathcal{E}$ satisfies pseudo-NTI. An SCF f is **robustly virtually implementable** in iteratively undominated strategies if and only if it is incentive compatible for every consistent type space and A-M measurable.*

Proof of Theorem 2: By Propositions 1 and 2, incentive compatibility for every consistent type space and A-M measurability are necessary conditions. We shall now establish that they are also sufficient, by constructing a canonical implementing mechanism. We note that the construction of the canonical mechanism of this section is a generalization of that in Theorem 1 once we take into account that the measurability algorithm may not stop at the first step.

In the mechanism $\tilde{\Gamma}$, every player i makes $(J+1)$ simultaneous announcements; in each the player announces an atom in the partition $\psi_i^* \in \Psi_i^*$:

$$M_i = M_i^0 \times M_i^1 \times \cdots \times M_i^J = \underbrace{\Psi_i^* \times \cdots \times \Psi_i^*}_{J+1}$$

for an integer J to be defined below. Correspondingly, the truthful s-th announce-

ment for type τ_i with pseudo-type t_i is $m_i^s = \Pi_i^*(t_i)$.

Define an SCF $x : T \to \Delta(A)$ by

$$x(t) = \frac{\alpha}{n} \sum_{i \in N} \sum_{h=0}^{L} \delta^h x_i^h[\Pi_i^h(t_i)](t) \ \ \forall t \in T$$

where $x_i^h[\Pi_i^h(t_i)]$ are arbitrary constant SCFs for $h = 0$, and are as constructed in Lemma 8 for each $h > 0$; $0 < \delta < 1$. Also, α is defined as

$$\alpha \equiv \frac{1}{1 + \delta + \delta^2 + \cdots + \delta^L}.$$

Note how x is A-M measurable by construction. Recall that, thanks to A-M measurability, we can abuse notation and write, for any $\psi \in \Psi^*$, $x(\psi) = x(t)$ whenever $\psi = \Pi^*(t)$.

Define the "bribe/punishment" lottery $\xi : N \times M \to \Delta(A)$ as follows:

$$\xi(i,m) = \begin{cases} \arg\min_{x(\psi_i, m_{-i}^0), \psi_i \in \Psi_i^*} \{V_i(x(\psi_i, m_{-i}^0) | m_i^0)\} & \text{if } \exists j \in \{1, \ldots, J\} \text{ s.t.} \\ & m_i^j \neq m_i^0 \text{ for } i \in N \text{ and} \\ & m^s = m^0 \ \forall \ s \in \{1, \ldots, j-1\}. \\ x(m^0) & \text{otherwise} \end{cases}$$

For any $i \in N$, define also

$$\ell_i(t) = x_i^L[\Pi_i^*(t_i)](t)$$

for any $t \in T$. Note that $x_i^L[\cdot]$ is A-M measurable.

Let the outcome function of the mechanism $\tilde{\Gamma}$ be $\tilde{g}$, defined as follows:

$$\tilde{g}(m) = \varepsilon x(m^0) + \frac{\varepsilon^2}{n} \sum_{i \in N} \xi(i,m) + \frac{(1 - \varepsilon - \varepsilon^2)}{J} \sum_{s=1}^{J} \tilde{f}(m^s),$$

where

$$\tilde{f}(m^s) = \frac{\varepsilon^2}{n} \sum_{i \in N} \ell_i(m_i^s, m_{-i}^0) + (1 - \varepsilon^2) f(m^s).$$

The next few paragraphs introduce several parameters, and fix their permissible values for the rest of the proof.

First, we can choose $\eta > 0$ small enough so that for every $h = 1, \ldots, L$,

$$\min_{i \in N, t_i \in T_i, \psi_i^h \neq \Pi_i^h(t_i)} V_i(x_i^h[\Pi_i^h(t_i)]|t_i) - V_i(x_i^h[\psi_i^h]|t_i) > 2\eta.$$

As in the proof of Theorem 1, $\eta > 0$ is well defined when the sets of pseudo-types are finite.

Second, for every SCF y, define

$$G_i(y) = \max_{t, t' \in T} |u_i(y(t'); \hat{\theta}(t)) - u_i(y(t_i, t'_{-i}); \hat{\theta}(t))|.$$

Choose $\delta > 0$ small enough so that for every $i \in N$ and every $h = 0, 1, \ldots, L-1$,

$$\eta > \sum_{i' \in N} \sum_{k=h+1}^{L} \delta^k G_i(x_{i'}^k).$$

Third, choose $\varepsilon > 0$ small enough so that for any $h = 1, \ldots, L$,

$$\begin{aligned} & \min_{i \in N, t_i \in T_i, \psi_i^h \neq \Pi_i^h(t_i)} \left\{ V_i(x_i^h[\Pi_i^h(t_i)]|t_i) - V_i(x_i^h[\psi_i^h]|t_i) - \eta \right\} \\ > \; & \frac{\varepsilon}{n} \max_{i \in N, t_i \in T_i, \psi_i^h \neq \Pi_i^h(t_i)} \left\{ V_i(x_i^h[\Pi_i^h(t_i)]|t_i) - V_i(x_i^h[\psi_i^h]|t_i) + \eta \right\}. \end{aligned}$$

And fourth, define $\eta_0(L), \eta_1(L) > 0$ with $\eta_0(L) < \eta_1(L)$ as follows:

$$\begin{aligned} \eta_0(L) &\equiv \varepsilon^2 \frac{\alpha \delta^L}{n^2} \min_{i \in N, t_i \in T_i, \psi_i^L \neq \Pi_i^L(t_i)} \left\{ V_i(x_i^L[\Pi_i^L(t_i)]|t_i) - V_i(x_i^L[\psi_i^L]|t_i) - \eta \right\} > 0; \\ \eta_1(L) &\equiv \varepsilon^2 \frac{\alpha \delta^L}{n^2} \max_{i \in N, t_i \in T_i, \psi_i^L \neq \Pi_i^L(t_i)} \left\{ V_i(x_i^L[\Pi_i^L(t_i)]|t_i) - V_i(x_i^L[\psi_i^L]|t_i) + \eta \right\} > 0. \end{aligned}$$

It is important to note that $\varepsilon, \eta, \delta, \eta_0(L)$, and $\eta_1(L)$ are chosen independently of the type space $\mathcal{T}$. Fix all of these variables at the specified levels.

The rest of the argument in the proof relies on two steps, as Claims 1.1 and 1.2 in Theorem 1, although it is somewhat more complicated. Specifically, the proof will require double use of mathematical induction. Claims 2.1 and 2.2 below, similar

to Claims 1.1 and 1.2 of Theorem 1, construct an induction step on the number of announcements j in the canonical mechanism for each agent. This serves to establish that if each agent i is using an iteratively undominated strategy, he must be reporting $\Pi_i^*(t_i)$ $(J+1)$ times when his pseudo-type is t_i. However, to establish Claim 2.1, a second induction argument is required, this time on h, the rounds of iteration in the A-M measurability algorithm. This is needed because the functions $x_i^h[\cdot]$ that are used to separate pseudo-types are not independent of the announcements made by others (unlike the ℓ_i's functions of Theorem 1). Now we proceed to complete the argument.

Fix an arbitrary consistent type space $\mathcal{T}$. All the analysis is invariant to the particular choice of type space made.

Claim 2.1: Let σ be an iteratively undominated strategy profile of the mechanism $\tilde{\Gamma}$. Then, for any $i \in N$, $\tau_i \in \mathcal{T}_i$, and $h = 0, 1, \ldots, L$, we have $\sigma_i^0(\tau_i) \subset \Pi_i^h(\hat{t}_i(\tau_i))$. In other words, $\sigma_i^0(\tau_i) = \Pi_i^*(\hat{t}_i(\tau_i))$ for any $\tau_i \in \mathcal{T}_i$ and $i \in N$.

Proof of Claim 2.1: We prove this step by induction with respect to h. Suppose $h = 0$. Then, $\Pi_i^0(\hat{t}_i(\tau_i)) = T_i$ for any $\tau_i \in \mathcal{T}_i$ and any $i \in N$. Therefore, the statement $\sigma_i^0(\tau_i) \subset \Pi_i^0(\hat{t}_i(\tau_i))$ in Claim 2.1 is trivially satisfied.

Suppose that $\sigma_i^0(\tau_i) \subset \Pi_i^h(\hat{t}_i(\tau_i))$ for any $\tau_i \in \mathcal{T}_i$ and any $h \leq L-1$. What we want to show is that $\sigma_i^0(\tau_i) \subset \Pi_i^L(\hat{t}_i(\tau_i))$, which equals $\Pi_i^*(\hat{t}_i(\tau_i))$, for any $\tau_i \in \mathcal{T}_i$ and any $i \in N$. Suppose, by way of contradiction, that there exists agent i of type τ_i for whom $\sigma_i^0(\tau_i) \subset \Pi_i^{L-1}(\hat{t}_i(\tau_i)) \backslash \Pi_i^L(\hat{t}_i(\tau_i))$. Consider agent i's strategy $\tilde{\sigma}_i$ with the following properties:

$$\begin{aligned} \sigma_i^j &= \tilde{\sigma}_i^j \ \forall j \geq 1, \\ \tilde{\sigma}_i^0(\tau_i^{'}) &= \sigma_i^0(\tau_i^{'}) \ \forall \tau_i^{'} \neq \tau_i \text{ and} \\ \tilde{\sigma}_i^0(\hat{t}_i(\tau_i)) &= \Pi_i^*(\hat{t}_i(\tau_i)). \end{aligned}$$

With Lemma 4 concerning the equivalence of types in mind, for any σ_{-i} under the inductive hypothesis, we have that the expected utility gain from the first term

of the outcome function is:

$$\begin{aligned} & \varepsilon\left\{U_i(x\circ(\tilde{\sigma}_i^0,\sigma_{-i}^0)|\tau_i)-U_i(x\circ\sigma^0|\tau_i)\right\} \\ = \;& \varepsilon\frac{\alpha\delta^L}{n}\left\{U_i(x_i^L[\Pi_i^L(\hat{t}_i(\tau_i))]\circ(\tilde{\sigma}_i^0,\sigma_{-i}^0)|\tau_i)-U_i(x_i^L[\Pi_i^L(\hat{t}_i(\tau_i))]\circ\sigma^0|\tau_i)\right\}. \end{aligned}$$

This is because no x_i^h, $h<L$, is affected by this strategy change and because for each $i^{'}\neq i$, $x_{i'}^L$ is measurable with respect to $\Psi_{i'}^L\times\Psi_{-i'}^{L-1}$ – recall that $\sigma_i^0(\tau_i)\subset\Pi_i^{L-1}(\hat{t}_i(\tau_i))\backslash\Pi_i^L(\hat{t}_i(\tau_i))$. Moreover, the latter expression we have just written is

$$\begin{aligned} \geq \;& \varepsilon\frac{\alpha\delta^L}{n}\left\{U_i(x_i^L[\Pi_i^L(\hat{t}_i(\tau_i))]\circ(\tilde{\sigma}_i^0,\sigma_{-i}^0)|\tau_i)-U_i(x_i^L[\Pi_i^L(\hat{t}_i(\tau_i))]\circ\sigma^0|\tau_i)-\sum_{i'\in N}\delta^L G_i(x_{i'}^L)\right\} \\ > \;& \varepsilon\frac{\alpha\delta^L}{n}\left\{U_i(x_i^L[\Pi_i^L(\hat{t}_i(\tau_i))]\circ(\tilde{\sigma}_i^0,\sigma_{-i}^0)|\tau_i)-U_i(x_i^L[\Pi_i^L(\hat{t}_i(\tau_i))]\circ\sigma^0|\tau_i)-\eta\right\} \\ > \;& \varepsilon^2\frac{\alpha\delta^L}{n^2}\left\{U_i(x_i^L[\Pi_i^L(\hat{t}_i(\tau_i))]\circ(\tilde{\sigma}_i^0,\sigma_{-i}^0)|\tau_i)-U_i(x_i^L[\Pi_i^L(\hat{t}_i(\tau_i))]\circ\sigma^0|\tau_i)+\eta\right\} \\ = \;& \eta_1(L). \end{aligned}$$

Thus, what agent i of type τ_i loses from the first term of the outcome function by misreporting in the 0-th announcement cannot be compensated by the "bribe/punishment" lottery, regardless of the other agents' announcements.

Hence, for any $\tau_i,\tau_i^{'}\in\mathcal{T}_i$ with $\hat{t}_i(\tau_i)=t_i$ and $\hat{t}_i(\tau_i^{'})=t_i^{'}$, $t_i\neq t_i^{'}$, we obtain

$$U_i(g\circ(\tilde{\sigma}_i,\sigma_{-i})|\tau_i)>U_i(g\circ\sigma|\tau_i).$$

The above inequality implies that player i will be strictly better off by telling the truth in the 0-th announcement, even if he misrepresents the rest of his announcements. Therefore, σ_i is strictly dominated by $\tilde{\sigma}_i$, which contradicts the hypothesis that σ is an iteratively undominated strategy profile. This completes the proof of Claim 2.1. ■

Claim 2.2: For every $i\in N$, let σ_i be an iteratively undominated strategy in the mechanism $\tilde{\Gamma}$. Suppose that $\sigma_i^s(\tau_i)=\Pi_i^*(\hat{t}_i(\tau_i))$ for all $i\in N,\ \tau_i\in\mathcal{T}_i$ and

$s \in \{0, \ldots, j\}$, where $0 \leq j \leq J-1$. Then

$$\sigma_i^{j+1}(\tau_i) = \Pi_i^*(\hat{t}_i(\tau_i)) \text{ for all } i \in N \text{ and all } \tau_i \in \mathcal{T}_i.$$

Proof of Claim 2.2: By Claim 2.1, we have proved that each agent tells the truth at the 0-th announcement. Thus, $\tilde{f}$ is strictly incentive compatible if f is incentive compatible.

Suppose, by way of contradiction, that $\sigma_i^{j+1}(\hat{t}_i(\tau_i)) \neq \Pi_i^*(\hat{t}_i(\tau_i))$ for some player i of some type $\tau_i \in \mathcal{T}_i$. So, by the very construction of the "bribe/punishment" lottery, he has to face the punishment. Define $\bar{\sigma}_i$ such that

$$\begin{aligned} \bar{\sigma}_i^s &= \sigma_i^s \ \forall s \neq j+1, \\ \bar{\sigma}_i^{j+1}(\tau_i^{'}) &= \sigma_i^{j+1}(\tau_i^{'}) \ \forall \tau_i^{'} \neq \tau_i, \\ \text{and } \ \bar{\sigma}_i^{j+1}(\tau_i) &= \Pi_i^*(\hat{t}_i(\tau_i)). \end{aligned}$$

Under the inductive hypothesis, if $\sigma_{-i}^{j+1}(\hat{t}_{-i}(\tau_{-i})) = \Pi_{-i}^*(\hat{t}_{-i}(\tau_{-i}))$ for all $\tau_{-i} \in \mathcal{T}_{-i}$, then by strict incentive compatibility of $\tilde{f}$ and by the definition of the "bribe/punishment" lottery $\xi(i,m)$, $\bar{\sigma}_i$ yields higher payoff than σ_i.

On the other hand, suppose that $\sigma_{i'}^{j+1}(\hat{t}_{i'}(\tau_{i'})) \in \mathcal{T}_{i'} \backslash \Pi_{i'}^*(\hat{t}_{i'}(\tau_{i'}))$ for some agent $i^{'} \neq i$ of some type $\tau_{i'} \in \mathcal{T}_{i'}$. Then, we choose J large enough so that

$$\eta_0(L) > \frac{1-\varepsilon-\varepsilon^2}{J}\gamma \geq \frac{1-\varepsilon-\varepsilon^2}{J}\left\{U_i(\tilde{f} \circ \sigma^{j+1}|\tau_i) - U_i(\tilde{f} \circ (\bar{\sigma}_i^j, \sigma_{-i}^{j+1})|\tau_i)\right\}.$$

Then, $\bar{\sigma}_i$ yields higher payoff than σ_i, which contradicts the hypothesis that σ_i is an iteratively undominated strategy of agent i. This completes the proof of Claim 2.2. ■

Claims 2.1 and 2.2 together show that there is a unique iteratively undominated strategy profile σ with the property that $\sigma_i^s(\hat{t}_i(\tau_i)) = \Pi_i^*(\hat{t}_i(\tau_i))$ for any $i \in N$, $\tau_i \in \mathcal{T}_i$, any consistent type space $\mathcal{T}$, and $s \in \{0, 1, \ldots, J\}$. The resulting outcome is

$$(1-\varepsilon^2)(1-\varepsilon-\varepsilon^2)f(\Pi^*(t)) + \varepsilon\left[(1-\varepsilon)(1+\varepsilon)^2 + \varepsilon\right]x(\Pi^*(t)).$$

Since the SCF f and x are A-M measurable, the resulting outcome is the same as

$$(1-\varepsilon^2)(1-\varepsilon-\varepsilon^2)f(t)+\varepsilon\left[(1-\varepsilon)(1+\varepsilon)^2+\varepsilon\right]x(t).$$

This is arbitrarily close to $f(t)$ for any $t \in T$ whenever $\varepsilon > 0$ is chosen small enough. This completes the proof of Theorem 2.■

7 The Relationship with Virtual Bayesian Implementation

All our results have been obtained using the very weak solution concept of iteratively undominated strategies. When robustness with respect to type spaces is a concern, it follows that there must be a connection with the approach that uses Bayesian equilibrium in every type space. This section explores this connection. First, consider the following definitions:

Let $\mathcal{B}(\Gamma)$ be the set of mixed-strategy Bayesian equilibria of the mechanism Γ.

Definition 13 (Robust Implementation in Bayesian Equilibrium) *An SCF f is said to be **robustly implementable** in mixed-strategy Bayesian equilibrium if there exists a mechanism $\Gamma = (M, g)$ such that $\mathcal{B}(\Gamma) \neq \emptyset$ and for any $\sigma^* \in \mathcal{B}(\Gamma)$, $g(\sigma^*(\tau)) = f(\hat{t}(\tau))$ for every $\tau \in \mathcal{T}$ and every consistent type space $\mathcal{T}$.*

Definition 14 (Robust Virtual Implementation in Bayesian Equilibrium) *An SCF f is **robustly virtually implementable** in mixed-strategy Bayesian equilibrium if, there exists $\bar{\varepsilon} > 0$ such that for any $\varepsilon \in (0, \bar{\varepsilon}]$, there exists an SCF f^ε for which $d(f, f^\varepsilon) < \varepsilon$ and f^ε is robustly implementable in mixed-strategy Bayesian equilibrium.*

Let us begin with our Theorem 1, which shows that the set of iteratively undominated strategies is not only unique but also *strict*. Thus, as an important by product, we obtain the following result for environments satisfying pseudo-TD.

Corollary 3 (Robust Virtual Bayesian Implementation) *Suppose an environment $\mathcal{E}$ satisfies pseudo-NTI and pseudo-TD. If an SCF is incentive compatible for*

every consistent type space $\mathcal{T}$, then it is robustly virtually implementable in mixed strategy Bayesian equilibrium.

Next, with the same argument, one can provide the following simple corollary to Theorem 2 if one does not assume pseudo-TD:

Corollary 4 ***(A Sufficient Condition for Robust Virtual Bayesian Implementation)*** *Suppose an environment $\mathcal{E}$ satisfies pseudo-NTI. An SCF f is robustly virtually implementable in mixed strategy Bayesian equilibrium if it is incentive compatible for any consistent type space and A-M measurable.*

It is important to note that A-M measurability is *not* necessary for robust virtual implementation in mixed strategy Bayesian equilibrium. To make this point, an elaboration of the example in Section 5 of Serrano and Vohra (2005) would suffice.[17] However, when the implementing mechanism is required to be *regular*, to be defined next, A-M measurability becomes necessary for robust virtual implementation in mixed strategy Bayesian equilibrium.

The next definitions are borrowed from Abreu and Matsushima (1992):

For every $i \in N$ and every partition Ψ_i, let $\Sigma_i(\Psi_i)$ denote the set of mixed strategies of player i that are measurable with respect to Ψ_i.

Definition 15 (pseudo-Bayesian Equilibrium) *The profile $\sigma \in \Sigma_1(\Psi_1) \times \cdots \times \Sigma_n(\Psi_n)$ is a **pseudo-Bayesian equilibrium** with respect to Ψ in Γ for a consistent type space $\mathcal{T}$ if for all $i \in N$ and all $\psi_i \in \Psi_i$, there exists **some** $\tau_i \in \mathcal{T}_i$ with $\hat{t}_i(\tau_i) \in \psi_i$ such that*

$$U_i(g \circ \sigma | \tau_i) \geq U_i(g \circ (\sigma_i^{'}, \sigma_{-i}) | \tau_i) \ \ \forall \sigma_i^{'} \in \Sigma_i$$

Definition 16 (Regular Mechanisms) *A mechanism Γ is said to be **regular** if for each Ψ there exists a pseudo-Bayesian equilibrium with respect to Ψ in Γ for any consistent type space.*

In particular, finite mechanisms – like the ones constructed in the proofs of Theorems 1 and 2 – are regular. Mechanisms that rely on the use of integer games – e.g., like the one constructed in Serrano and Vohra (2005) – are not regular.

[17] Although Serrano and Vohra (2005) restricts attention to implementation in pure strategies, the argument can be extended to also cover mixed strategies.

The next result extends a result in Abreu and Matsushima (1992) to our settings:

Proposition 3 *If an SCF is robustly virtually implementable in mixed strategy Bayesian equilibrium by a* ***regular mechanism****, then it is A-M measurable.*

Proof of Proposition 3: Since f is robustly virtually implementable in Bayesian equilibrium, there exists f^{ε} that is exactly implementable in Bayesian equilibrium and $d(f, f^{\varepsilon}) < \varepsilon$ for $\varepsilon > 0$ sufficiently small for any consistent type space. Consider a "regular" mechanism $\Gamma = (M, g)$ that exactly implements the SCF f^{ε} in mixed Bayesian equilibrium for any consistent type space. Fix an arbitrary consistent type space $\mathcal{T}$. Let $\sigma \in \times_{i \in N} \Sigma_i(\Psi_i^*)$ be a pseudo-Bayesian equilibrium with respect to Ψ^*. Note that σ is measurable with respect to Ψ^*. What we want to show here is that σ is a Bayesian equilibrium as well.

If $m_i = \sigma_i(\tau_i)$ is a best response for player i of type τ_i, then m_i is also a best response for player i of any type $\tau_i^{'}$ such that $\hat{t}_i(\tau_i^{'}) \in \rho_i(t_i, \Psi_{-i}^*)$. That is, this implies that for any $\psi_i \in \Psi_i^*$, for any $\tau_i, \tau_i^{'} \in \mathcal{T}_i$ with $\hat{t}_i(\tau_i), \hat{t}_i(\tau_i^{'}) \in \psi_i$, the best responses of player i of type τ_i and $\tau_i^{'}$ to any σ_{-i} that is measurable with respect to Ψ_{-i}^* are the same. Then, it follows that any pseudo-Bayesian equilibrium σ that is measurable with respect to Ψ^* is in fact a Bayesian equilibrium. Since $f^{\varepsilon} = g \circ \sigma$ by our hypothesis that f^{ε} is exactly implementable in Bayesian equilibrium, f^{ε} is measurable with respect to Ψ^* and therefore it must be A-M measurable. Finally, for a sufficiently small $\varepsilon > 0$, it follows that f is A-M measurable if and only if f^{ε} is A-M measurable. ■

Putting together this proposition and Theorem 2, we arrive at the following:

Corollary 5 ***(A Characterization of Robust Virtual Bayesian Implementation)*** *Suppose an environment $\mathcal{E}$ satisfies pseudo-NTI. An SCF f is robustly virtually implementable in mixed strategy Bayesian equilibrium by a regular mechanism if and only if it is incentive compatible for any consistent type space and A-M measurable.*

On the other hand, the usual approach for a fixed type space to (exact and virtual) Bayesian implementation has ruled out the consideration of mixed strategies.[18] We show next that if one includes robustness considerations with respect to type spaces,

[18]Duggan (1997) is a notable exception.

the distinction between pure and mixed strategy equilibrium implementation is of no significance:

Proposition 4 *An SCF is robustly virtually implementable in pure-strategy Bayesian equilibrium if and only if it is robustly virtually implementable in mixed-strategy Bayesian equilibrium.*

Proof of Proposition 4: That full implementation in mixed strategy equilibrium implies full implementation in pure equilibrium is obvious. We argue the opposite direction.

Suppose not. There exists an SCF f that is robustly virtually implementable in pure Bayesian equilibrium that is not robustly virtually implementable in mixed equilibrium. This means that any mechanism that virtually implements f in pure equilibrium over every consistent type space has an equilibrium in properly mixed strategies whose outcome does not approximate f. But then, one can construct a sufficiently large consistent type space and perform a purification of that equilibrium. The result is a pure-strategy Bayesian equilibrium of the mechanism whose outcome is far from f. This contradicts that f is robustly virtually implementable in pure-strategy equilibrium. ■

Thus, while implementation in pure or mixed equilibrium may give different answers for a fixed type space, that difference goes away when one requires robustness in implementation with respect to type spaces.

8 Conclusion

By proposing a reinterpretation of the Wilson doctrine – mechanisms be allowed to depend on first-order beliefs, besides payoff types – we have shown that robust virtual implementation in iteratively undominated strategies is "almost always" as powerful as it can possibly be. Indeed, the limits of implementation are given by incentive compatibility, but every incentive compatible SCF can be robustly virtually implemented. Thus, even if one insists on robustness of implementation results with respect to type spaces, there is a significant gap between the very restrictive results offered by exact implementation and the much more permissive ones offered by the virtual approach.

References

Abreu, D. and H. Matsushima (1992): Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information, Unpublished Manuscript, Princeton University.

Abreu, D. and A. Sen (1991): Virtual Implementation in Nash Equilibrium, Econometrica, 59, 997-1021.

Battigalli, P. and M. Siniscalchi (2003): Rationalisation and Incomplete Information, Advances in Theoretical Economics, 3.

Bergemann, D. and S. Morris (2005a): Robust Mechanism Design, Econometrica, 73, 1771-1813.

Bergemann, D. and S. Morris (2005b): Robust Implementation: The Role of Large Type Spaces, Unpublished Manuscript, Cowles Foundation, Yale University.

Bergemann, D. and S. Morris (2007): Strategic Distinguishability with an Application to Robust Virtual Implementation, Unpublished Manuscript, Yale University and Princeton University.

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

Brandenburger, A. and E. Dekel (1993): Hierarchies of Beliefs and Common Knowledge, Journal of Economic Theory, 59, 189-198.

Chakravorti, B. (1992): Efficiency and Mechanisms with no Regret, International Economic Review, 33, 45-59.

Duggan, J. (1997): Virtual Bayesian Implementation, Econometrica, 65, 1175-1199.

Jackson, M. O. (1991): Bayesian Implementation, Econometrica, 59, 461-477.

Jackson, M. O. (2001): A Crash Course in Implementation Theory, Social Choice and Welfare, 18, 655-708.

Jehiel, P., M. Meyer-ter-Vehn, B. Moldovanu and B. Zame (2006): The Limits of Ex Post Implementation, Econometrica, 74, 585-610.

Maskin, E. S. and T. Sjostrom (2002): Implementation Theory, in Handbook of Social Choice and Welfare (vol. I), ed. by K. J. Arrow, A. Sen and K. Suzumura, New York, Elsevier Science B.V.

Mertens, J-F and S. Zamir (1985): Formulation of Bayesian Analysis for Games with Incomplete Information, International Journal of Game Theory, 14, 1-29.

Myerson, R. B. (1989): Mechanism Design, in The New Palgrave: Allocation, Information, and Markets, ed. by J. Eatwell, M. Milgate and P. Newman, Norton, New York.

Palfrey, T. R. (2002): Implementation Theory, in Handbook of Game Theory with Economic Applications (vol. III), ed. by R. J. Aumann and S. Hart, New York, Elsevier Science.

Palfrey, T. R. and S. Srivastava (1987): On Bayesian Implementable Allocations, Review of Economic Studies, 54, 193-208.

Palfrey, T. R. and S. Srivastava (1989): Implementation with Incomplete Information in Exchange Economies, Econometrica, 57, 115-134.

Palfrey, T. R. and S. Srivastava (1993): Bayesian Implementation, Harwood Academic Publishers, New York.

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

Serrano, R. (2004): The Theory of Implementation of Social Choice Rules, SIAM Review, 46, 377-414.

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

Serrano, R. and R. Vohra (2005): A Characterization of Virtual Bayesian Implementation, Games and Economic Behavior, 50, 312-331.

Wilson, R. (1987): Game Theoretic Analysis of Trading Processes, in Advances in Economic Theory, ed. by T. Bewley, Cambridge University Press.