9-2023

Interim rationalizable implementation of functions

Takashi KUNIMOTO
Singapore Management University, tkunimoto@smu.edu.sg

Rene SARAN

Roberto SERRANO

Citation

KUNIMOTO, Takashi; SARAN, Rene; and SERRANO, Roberto. Interim rationalizable implementation of functions. (2023). *Mathematics of Operations Research*. 1-34.
Available at: https://ink.library.smu.edu.sg/soe_research/2689

Interim Rationalizable Implementation of Functions

Takashi Kunimoto,[a,*] **Rene Saran,**[b] **Roberto Serrano**[c]

[a] School of Economics, Singapore Management University, Singapore 178930; [b] Department of Economics, University of Cincinnati, Cincinnati, Ohio 45221; [c] Department of Economics, Brown University, Providence, Rhode Island 02912

*Corresponding author

Contact: tkunimoto@smu.edu.sg, https://orcid.org/0000-0002-7798-7435 (TK); rene.saran@uc.edu (ReS); roberto_serrano@brown.edu (RoS)

Received: April 9, 2021
Revised: July 25, 2022; April 4, 2023
Accepted: July 30, 2023
Published Online in Articles in Advance: September 13, 2023

MSC2020 Subject Classifications: Primary: 91B03, 91A27; secondary: 91B26

https://doi.org/10.1287/moor.2022.0202

Abstract. This paper investigates rationalizable implementation of social choice functions (SCFs) in incomplete information environments. We identify weak interim rationalizable monotonicity (weak IRM) as a novel condition and show it to be a necessary and almost sufficient condition for rationalizable implementation. We show by means of robust examples that interim rationalizable monotonicity (IRM), found in the literature, is strictly stronger than weak IRM and that IRM is not necessary for rationalizable implementation, as had been previously claimed. These examples also demonstrate that Bayesian monotonicity, the key condition for full Bayesian implementation, is not necessary for rationalizable implementation. That is, rationalizable implementation can be more permissive than Bayesian implementation. We revisit well-studied classes of economic environments and show that the SCFs considered there are interim rationalizable implementable. A comprehensive discussion of related issues, including well-behaved mechanisms, mechanisms satisfying the best response property, double implementation, and responsive SCFs is also provided.

Funding: This work was supported by Ministry of Education, Singapore [Grant MOE Academic Research Fund Tier 2/MOE-T2EP402A20-0].

Keywords: Bayesian incentive compatibility • Bayesian monotonicity • weak interim rationalizable monotonicity • interim rationalizable monotonicity • implementation • rationalizability

1. Introduction

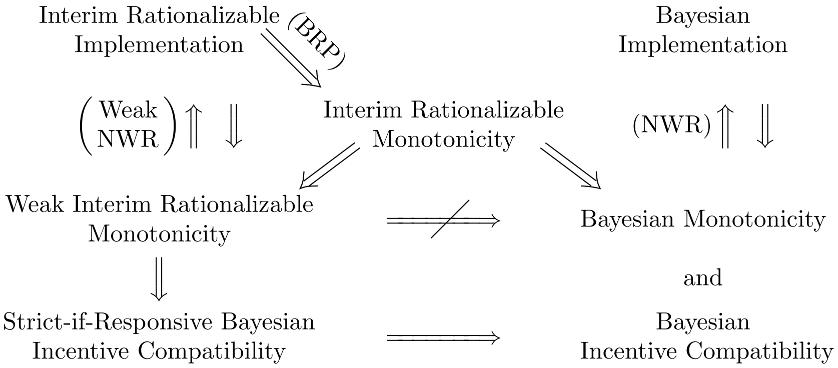
A leading solution concept in game theory is rationalizability (Bernheim [12], Pearce [35], Brandenburger and Dekel [13], Lipman [25]). When players are rational and there is common belief among them that this is the case, they play their rationalizable strategies, without necessarily imposing the additional assumption that their beliefs are correct, as is the case in an equilibrium.[1] Its extension to Bayesian games of incomplete information, our concern in this paper, is the notion of interim correlated rationalizability, from Dekel et al. [16], which will be defined in a later section.[2]

Despite the impressive effort made by implementation theorists in the 1980s and 1990s, using a plethora of game-theoretic solution concepts, a characterization of the rules that are implementable in rationalizable strategies under incomplete information has remained an open problem. The current paper settles this issue, by essentially providing such a characterization, for the case of single-valued rules or social choice functions (SCFs). A previous working paper (Bergemann and Morris [7]) provides valuable results for the case of finite mechanisms.[3]

Our main finding is to propose a novel condition, which we term weak interim rationalizable monotonicity (weak IRM), that is necessary and almost sufficient for implementation in interim rationalizable strategies (Theorems 1 and 2). Weak IRM is a weakening of the IRM condition proposed in Bergemann and Morris [7], which will be shown not to be necessary for rationalizable implementation (Section 7).[4] We stress this point because Oury and Tercieux [33] make an incorrect claim that IRM is necessary for interim rationalizable implementation in its footnote 4. IRM, but not weak IRM, implies Bayesian monotonicity (Bergemann and Morris [7]), a necessary condition for implementation in Bayesian equilibrium.[5] Indeed, we show in several examples in Section 7 that weak IRM can be satisfied even when Bayesian monotonicity fails. Our results thus demonstrate that rationalizable implementation may be more permissive than equilibrium implementation. Figure 1 summarizes the relationship between Bayesian implementation and interim rationalizable implementation, which is elaborated in detail later (Sections 5, 7, and 8).

The finding just described, that making the assumption of equilibrium or correct expectations may be restricting the set of rules that can be decentralized by means of play in mechanisms, ought to be compared with results in complete information environments. In contrast to our finding, Bergemann et al. [11] and Xiong [45] show that

Figure 1. Relationship between interim rationalizable implementation and Bayesian implementation.


Interim Rationalizable Implementation
(BRP)
Bayesian Implementation
Weak NWR
Interim Rationalizable Monotonicity
(NWR)
Weak Interim Rationalizable Monotonicity
Bayesian Monotonicity
and
Strict-if-Responsive Bayesian Incentive Compatibility
Bayesian Incentive Compatibility

Note. BRP, best response property; NWR, no-worst-rule.

rationalizable implementation of SCFs under complete information is more restrictive than equilibrium implementation. For set-valued rules, however, Kunimoto and Serrano [24] come to the reverse conclusion that rationalizable implementation is generally more permissive than equilibrium implementation under complete information.[6] For general correspondences, Kunimoto and Serrano [24] identify uniform monotonicity, which is a weakening of the classic Maskin monotonicity (Maskin [26]) and that reduces to it in the case of SCFs as a necessary and almost sufficient condition for rationalizable implementation. Because Maskin monotonicity is necessary and almost sufficient for Nash implementation, regardless of whether one wishes to implement SCFs or general correspondences, finding rules that are Nash implementable but not implementable in rationalizable strategies is generally very difficult: such rules are Maskin monotonic, which in addition to the other weak conditions identified in Kunimoto and Serrano [24], will also make them rationalizably implementable. Conversely, it is easy to find set-valued rules that are implementable in rationalizable strategies but not in Nash equilibrium.

This paper's results show that the permissiveness of rationalizable implementation, in comparison with equilibrium implementation, carries over to incomplete information environments, but now even for SCFs.[7] This happens if the implementing mechanism in rationalizable strategies fails to have equilibria, showcasing the additional requirement of the best-response correspondence having fixed points (see Examples 2 and 6 in Section 7 for this). Note how one should not discard a mechanism outright because it fails to have equilibria; if one accepts more flexible patterns of behavior, such as different bounded-rationality rules of thumb or rationalizability, which concerns us here, one should be open to that possibility. Certainly, to understand the restrictions that rationalizable play imposes on implementability, such mechanisms must be considered, as we do here. In doing so, we construct an implementing mechanism that works using rationalizable strategies, regardless of whether it has equilibria or not. In the quest of the general conditions for rationalizable implementation, we view this as an improvement over the canonical constructions in the existing literature. We plan to explore generalizations of the findings in Kunimoto and Serrano [24], as well as those in the current study, in a separate paper posing the question of set-valued rules under incomplete information.[8]

The plan of the paper is as follows. Section 2 presents preliminaries. Section 3 introduces our notion of implementation in interim rationalizable strategies. Weak IRM, as the necessary condition for interim rationalizable implementation, is presented in Section 4. Section 5 relates weak IRM and IRM to previous conditions (Bayesian incentive compatibility and Bayesian monotonicity). Section 6 shows that weak IRM and an additional weak condition are sufficient for interim rationalizable implementation. Section 7 demonstrates the significance of our results in a series of well-studied classes of economic environments and, in particular, shows that IRM and Bayesian monotonicity are *not* necessary for interim rationalizable implementation. Section 8 discusses a number of extensions of our results and Section 9 concludes the paper. Most proofs are relegated to the appendices.

2. Preliminaries

Let $I = \{1, \ldots, n\}$ denote the finite set of agents or players, and T_i be a finite set of types of agent i. Let $T \equiv T_1 \times \cdots \times T_n$, and $T_{-i} \equiv T_1 \times \cdots \times T_{i-1} \times T_{i+1} \times \cdots \times T_n$.[9] Let $\Delta(T_{-i})$ denote the set of probability distributions over T_{-i}. Each agent i has a system of "interim" beliefs that is expressed as a function $\pi_i : T_i \to \Delta(T_{-i})$. Then, we call $(T_i, \pi_i)_{i\in I}$ a *type space*. Let A denote a countable set of pure outcomes, which are assumed to be independent of the information state. Let $\Delta(A)$ be the set of probability distributions over A. Agent i's state dependent von Neumann-Morgenstern utility function is denoted $u_i : \Delta(A) \times T \to \mathbb{R}$. We can now define an *environment* as $\mathcal{E} = (A, \{u_i, T_i, \pi_i\}_{i\in I})$.

A (stochastic) *social choice function* (SCF) is a single-valued function $f : T \to \Delta(A)$.[10] In any arbitrary type space, some states might be deemed impossible by all agents, complete information environments being the most notable case. If the designer thinks that allocations in states occurring with probability zero are irrelevant, the designer's interest in implementation applies only to a subset of T. To take this into account, we let $T^* \subseteq T$ be the set of states that the designer cares about. We assume that T^* is such that

$$\{t \in T : \exists i \in I \text{ s.t. } \pi_i(t_i)[t_{-i}] > 0\} \subseteq T^*.$$

This formulation requires the designer to care about all states that are deemed possible by at least one agent while at the same time it still allows the designer's flexibility to care about states that are deemed impossible by all agents.[11] Consider any two SCFs f, f'. We say that f and f' are *equivalent* (denoted by $f \approx f'$) if $f(t) = f'(t)$ for all $t \in T^*$.[12]

A *mechanism* (or *game form*) $\Gamma = ((M_i)_{i \in I}, g)$ describes: (i) a nonempty countable message space M_i for each agent i and (ii) an outcome function $g : M \to \Delta(A)$, where $M = \prod_{i \in I} M_i$. Let $\Gamma^{DR} = ((T_i)_{i \in I}, f)$ denote the *direct revelation mechanism* (or direct mechanism) associated with an SCF f, that is, a mechanism where $M_i = T_i$ for all i and $g = f$.

In the direct mechanism associated with an SCF f, the interim expected utility of agent i of type t_i who pretends to be of type t_i', whereas all other agents truthfully announce their types, is defined as

$$U_i(f; t_i' | t_i) \equiv \sum_{t_{-i} \in T_{-i}} \pi_i(t_i)[t_{-i}] u_i\big(f(t_i', t_{-i}), (t_i, t_{-i})\big).$$

Let $U_i(f | t_i) = U_i(f; t_i | t_i)$.

For any $i \in I$ and function $y : T_{-i} \to \Delta(A)$, we define

$$U_i(y | t_i) \equiv \sum_{t_{-i} \in T_{-i}} \pi_i(t_i)[t_{-i}] u_i\big(y(t_{-i}), (t_i, t_{-i})\big).$$

3. Implementation in Interim Rationalizable Strategies

We adopt *interim correlated rationalizability* (Dekel et al. [16]) as a solution concept and investigate the implications of implementation in interim correlated rationalizable strategies.[13] We fix a mechanism $\Gamma = ((M)_{i \in I}, g)$ and define a message correspondence profile $S = (S_1, \ldots, S_n)$, where each $S_i : T_i \to 2^{M_i}$, and we write $\mathcal{S}$ for the collection of message correspondence profiles. The collection $\mathcal{S}$ is a lattice with the natural ordering of set inclusion: $S \leq S'$ if $S_i(t_i) \subseteq S_i'(t_i)$ for all $i \in I$ and $t_i \in T_i$. The largest element is $\overline{S} = (\overline{S}_1, \ldots, \overline{S}_n)$, where $\overline{S}_i(t_i) = M_i$ for each $i \in I$ and $t_i \in T_i$. The smallest element is $\underline{S} = (\underline{S}_1, \ldots, \underline{S}_n)$, where $\underline{S}_i(t_i) = \emptyset$ for each $i \in I$ and $t_i \in T_i$.

We define an operator b to iteratively eliminate never best responses. The operator $b : \mathcal{S} \to \mathcal{S}$ is thus defined as: for every $i \in I$ and $t_i \in T_i$,

$$b_i(S)[t_i] \equiv \left\{ m_i : \begin{array}{l} \exists \lambda_i \in \Delta(T_{-i} \times M_{-i}) \text{ such that} \\ (1)\, \lambda_i(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}(t_{-i}); \\ (2)\, \operatorname{marg}_{T_{-i}} \lambda_i = \pi_i(t_i); \\ (3)\, m_i \in \arg\max\limits_{m_i'} \sum\limits_{t_{-i}, m_{-i}} \lambda_i(t_{-i}, m_{-i}) u_i(g(m_i', m_{-i}), (t_i, t_{-i})) \end{array} \right\}.$$

Observe that b is increasing by definition: that is, $S \leq S' \Rightarrow b(S) \leq b(S')$. By Tarski's fixed-point theorem, there is a largest fixed point of b, which we label $S^{\Gamma(T)}$. Thus, (i) $b(S^{\Gamma(T)}) = S^{\Gamma(T)}$ and (ii) $b(S) = S \Rightarrow S \leq S^{\Gamma(T)}$.

We can also construct the fixed point $S^{\Gamma(T)}$ by starting with $\overline{S}$, the largest element of the lattice, and iteratively applying the operator b. Let the message correspondence profile $S^{\Gamma(T),0} = \overline{S}$ and, for all $i \in I$, $t_i \in T_i$, $k \geq 1$, iteratively define,

$$S_i^{\Gamma(T),k}(t_i) \equiv b_i\big(S^{\Gamma(T),k-1}\big)[t_i].$$

If the message sets are finite, we have

$$S_i^{\Gamma(T)}(t_i) \equiv \bigcap_{k \geq 0} S_i^{\Gamma(T),k}(t_i)$$

for each $i \in I$ and $t_i \in T_i$. However, because the mechanism Γ may be infinite, transfinite induction may be necessary to reach the fixed point. Thus, $S_i^{\Gamma(T)}(t_i)$ are the sets of messages surviving (transfinite) iterated deletion of never best responses of type t_i of agent i.[14] We denote by σ_i a selection from $S_i^{\Gamma(T)}$ and call it a rationalizable strategy of agent i. We recall the following structure of $S^{\Gamma(T)}$:

$$S^{\Gamma(T)} = \prod_{i \in I} S_i^{\Gamma(T)}.$$

Definition 1. A mechanism Γ implements an SCF f in interim rationalizable strategies if there exists an SCF $\hat{f} \approx f$ such that the following two conditions hold:

1. Nonemptiness: $S_i^{\Gamma(T)}(t_i) \neq \emptyset$ for all $t_i \in T_i$ and $i \in I$.
2. Uniqueness: for any $t \in T$, $m \in S^{\Gamma(T)}(t)$ implies $g(m) = \hat{f}(t)$.

Remark. The uniqueness requirement in interim rationalizable implementation is stronger than the usual one, because we require that every rationalizable strategy profile induces outcomes specified by the equivalent SCF $\hat{f}$ over the entire T rather than T^*. This strengthening allows us to obtain a clean characterization for interim rationalizable implementation, whereas it makes our notion more stringent than the one used by Bergemann et al. [11] and Xiong [45], which can ignore many zero-probability states in complete information environments.

We say that an SCF f is *implementable in interim rationalizable strategies* if there exists a mechanism Γ that implements f in interim rationalizable strategies.

4. Necessity for Implementation of an SCF in Interim Rationalizable Strategies

In this section, we uncover a necessary condition for interim rationalizable implementation of an SCF. First, we turn to some preliminary definitions.

Definition 2. A *deception* is a profile of correspondences $\beta = (\beta_1, \ldots, \beta_n)$ such that $\beta_i : T_i \to 2^{T_i}$ and $t_i \in \beta_i(t_i)$ for all $t_i \in T_i$ and $i \in I$.

Intuitively, $\beta_i(t_i)$ describes the set of possible types which agent i of type t_i pretends to be in the direct mechanism associated with the SCF. This is an important consideration in implementation theory because the designer aims to elicit the types of the agents, typically by letting them announce their types, and the designer is fully aware that the agents can announce any types as long as it is in their interest to do so.

Remark. These set-valued deceptions have already been used in previous literature on interim rationalizable implementation (Bergemenn and Morris [7], Oury and Tercieux [33]). The requirement that $t_i \in \beta_i(t_i)$ for all t_i is made to simplify the writing of some steps in the proof below. It is not essential at all for our results.

Definition 3. A deception β is *acceptable for an SCF f* if, for all $t, t' \in T$, $t' \in \beta(t) \Rightarrow f(t) = f(t')$; otherwise, β is *unacceptable for f*.

Unacceptable deceptions are a concern for the designer since the agents undermine the designer's goal of implementing the outcome $f(t)$ for any $t \in T$.

Given an SCF f, for each $i \in I$ and $t_i \in T_i$, define

$$Y_i[t_i, f] \equiv \left\{ y : T_{-i} \to \Delta(A) : \begin{array}{ll} \text{either} & y(t_{-i}) = f(t_i, t_{-i}),\ \forall t_{-i} \in T_{-i} \\ \text{or} & U_i(f | t_i) > U_i(y | t_i) \end{array} \right\}.$$

Thus, $Y_i[t_i, f]$ is the collection of all mappings $y : T_{-i} \to \Delta(A)$ that individual i of type t_i considers to be "equivalent" to f or strictly worse than f.

For any SCF f and individual $i \in I$, we define a binary relation $\sim_i^f$ on $T_i \times T_i$ as follows: We say that $t_i \sim_i^f t_i'$ if f is not responsive to this change in i's type, that is,

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

Otherwise, we say $t_i \nsim_i^f t_i'$. Notice that $\sim_i^f$ is symmetric, that is, $t_i \sim_i^f t_i'$ if and only if $t_i' \sim_i^f t_i$. We say that an SCF f is *nonresponsive to agent i's type* if $t_i \sim_i^f t_i'$ for all $t_i, t_i' \in T_i$.

Definition 4. A deception β that is unacceptable for an SCF f is *weakly refutable* if there exist $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \nsim_i^f t_i$ such that for all $\psi_i \in \Delta(T_{-i} \times T)$ satisfying $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$, there exists an SCF f' such that $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$ and

$$\sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i\big(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})\big).$$

Unlike equilibrium, the solution concept of rationalizability allows different types of an agent to hold distinct beliefs about the behavior of the other agents. To illustrate this while keeping matters simple, suppose for each type $\tilde{t}_j$ of each agent j we can find a strategy profile $\sigma_{-j}^{\tilde{t}_j}$ such that $\sigma_{-j}^{\tilde{t}_j}(t_{-j}) \in S_{-j}^{\Gamma(T)}(t_{-j})$, for all t_{-j}, which rationalizes the behavior of type $\tilde{t}_j$ (i.e., type $\tilde{t}_j$ has a rationalizable message that is a best response to the belief that the other agents play according to the rationalizable strategy profile $\sigma_{-j}^{\tilde{t}_j}$). Now suppose that instead of reporting their own rationalizable messages, agents use the deception β (i.e., agents of types $\hat{t}$ report rationalizable messages corresponding to types in $\beta(\hat{t})$). When the deception β is weakly refutable, the designer finds an agent's type (say, type t_i of agent i) as an ally to undermine the deception. Specifically, this type finds a collection of SCFs, one for each belief $\psi_i \in \Delta(T_{-i} \times T)$ that is compatible with the fact that the other agents are using the deception β_{-i}. Notice that the belief ψ_i is defined over $T_{-i} \times T$ rather than $T_{-i} \times T_{-i}$ because player i is aware that types $\hat{t}_{-i}$ are playing messages that are rationalizable for types $\beta_{-i}(\hat{t}_{-i})$, which in turn rationalize the behavior of different types of player i. Therefore, the rationalizable messages for types $\beta_{-i}(\hat{t}_{-i})$ could vary depending on which type of player i's behavior those of types $\beta_{-i}(\hat{t}_{-i})$ rationalize. For instance, $\sigma_{-i}^{t_i}(\beta_{-i}(\hat{t}_{-i})) \in S_{-i}^{\Gamma(T)}(\beta_{-i}(\hat{t}_{-i}))$ that rationalize the behavior of type t_i of player i might be different from $\sigma_{-i}^{t_i'}(\beta_{-i}(\hat{t}_{-i})) \in S_{-i}^{\Gamma(T)}(\beta_{-i}(\hat{t}_{-i}))$ that rationalize the behavior of type t_i' of player i. Thus, when contemplating the behavior of types $\hat{t}_{-i}$ under the deception β, player i needs to form a belief over messages in $\cup_{\tilde{t}_i \in T_i}\{\sigma_{-i}^{\tilde{t}_i}(\beta_{-i}(\hat{t}_{-i}))\}$, which explains why the domain of ψ_i includes T_i as a component.

It is instructive to appreciate this feature of ψ_i in comparison with equilibrium implementation in incomplete information environments. For instance, in Bayesian implementation, all players share a common belief that one particular equilibrium strategy profile σ^* is played in the mechanism. Then, when contemplating the behavior of types $\hat{t}_{-i}$ under the deception β, player i's belief is simply that types $\hat{t}_{-i}$ report $\sigma_{-i}^*(\beta_{-i}(\hat{t}_{-i}))$, which is independent of player i's type.

The collection of SCFs that the ally finds to undermine the deception is required to satisfy the following two properties. First, by definition, each type $\tilde{t}_i$ places each of these SCFs f' in the strictly lower contour set of f under truth-telling whenever $f'(\tilde{t}_i, \cdot) \neq f(\tilde{t}_i, \cdot)$. Second, when the deception β is used, then under belief ψ_i, type t_i strictly prefers the corresponding SCF f' in the collection to f. If one insists on restricting the collection of SCFs to those f' that are nonresponsive to agent i's type, then one would speak of *strong* refutability. Under this restriction, there is a mapping $y : T_{-i} \to \Delta(A)$ such that $f'(\tilde{t}_i, \cdot) = y$ for all $\tilde{t}_i$. Then, the requirement that $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$ means that $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$. As discussed in the next section, this will be important to understand the difference with the previous condition proposed in the literature.

Definition 5. An SCF f satisfies *weak interim rationalizable monotonicity (weak IRM)* if every deception β that is unacceptable for f is weakly refutable.

If an SCF satisfies weak IRM, the designer can plan on using the services of the ally identified in the definition of weak refutability in order to succeed in the designer's attempt of implementing f.[15] If the designer insisted on the deception being "strongly" refutable, as defined in the next section, then the SCF would satisfy IRM, a stronger condition introduced in the literature (Bergemann and Morris [7], Oury and Tercieux [33]). In particular, it is claimed in Oury and Tercieux ([33], footnote 4) that IRM is necessary for the interim rationalizable implementation of SCFs. We will show this claim to be incorrect in the sequel.

Next, we present our first main result, which shows that weak IRM is necessary for implementation in rationalizable strategies.

Theorem 1. *If an SCF f is implementable in interim rationalizable strategies, then there exists an SCF $\hat{f} \approx f$ that satisfies weak IRM.*

5. Weak IRM, IRM, and Other Relevant Conditions

In this section, we investigate the connections between weak IRM, IRM, and the conditions of incentive compatibility and Bayesian monotonicity, central in the characterization of SCFs that are implementable in Bayesian equilibrium. Further connections will be uncovered in a later section, after we state and prove our sufficiency result.

Definition 6. An SCF f satisfies *Bayesian incentive compatibility (BIC)* if for all $i \in I$ and $t_i \in T_i$,

$$U_i(f \mid t_i) \geq U_i(f; t_i' \mid t_i), \ \forall t_i' \in T_i.$$

If these constraints are strict whenever $t_i \nsim_i^f t_i'$, then we say that f satisfies *strict-if-responsive Bayesian incentive compatibility (SIRBIC)*.

Clearly, SIRBIC is a strenghthening of BIC, whereas it is a weakening of strict IC, which imposes strict inequalities on all incentive constraints. Then, we can show the following.

Lemma 1. *If an SCF f satisfies weak IRM, then it satisfies SIRBIC.*

Remark. It is well known that BIC is necessary for implementability in Bayesian equilibrium. Lemma 1, along with Theorem 1, shows that SIRBIC, a stronger condition, is necessary for implementation in interim rationalizable strategies.

As discussed in the previous section when we defined weak refutability, one can propose its stronger version.

Definition 7. A deception β that is unacceptable for an SCF f is *strongly refutable* if there exist $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \not\sim_i^f t_i$ such that for all $\psi_i \in \Delta(T_{-i} \times T)$ satisfying $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$, there exists an SCF f' such that f' is nonresponsive to agent i's type, $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$, and

$$\sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i\big(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})\big).$$

Remark. The SCF f' in the statement for strong refutability is required to be nonresponsive to agent i's type, as opposed to allowing f' that could respond to a change in agent i's type in the statement for weak refutability. This additional requirement for strong refutability, in conjunction with the stipulation that $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$, implies that there exists a mapping $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ that is strictly preferred to f by type t_i of agent i when the deception β is used. Interim rationalizable monotonicity introduced by Bergemann and Morris [7] requires the existence of such a mapping $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ to undermine an unacceptable deception.[16] Indeed, as we show next, interim rationalizable monotonicity is equivalent to strong refutability of every unacceptable deception.

Definition 8. An SCF f satisfies *IRM* if, for every deception β that is unacceptable for f, there exist $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \not\sim_i^f t_i$ such that for all $\phi_i \in \Delta(T_{-i} \times T_{-i})$ satisfying $\phi_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$, there exists $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such that

$$\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i\big(y(\tilde{t}_{-i}), (t_i, t_{-i})\big) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i\big(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})\big).$$

Lemma 2. *An SCF f satisfies IRM if and only if every deception β that is unacceptable for f is strongly refutable.*

As it is clear that strong refutability implies its weak version, we state the following result without proof.

Corollary 1. *If an SCF f satisfies IRM, it also satisfies weak IRM.*

A *single-valued deception* β^s is a profile of functions $(\beta_1^s, \ldots, \beta_n^s)$ such that $\beta_i^s : T_i \to T_i$ for all $i \in I$. The single-valued deception β^s is *acceptable for an SCF f* if, for any $t \in T$, $f(\beta^s(t)) = f(t)$; otherwise, β^s is *unacceptable for f*.

It is worth distinguishing between single-valued deceptions and our set-valued deceptions that happen to be single valued. Recall that our definition of a set-valued deception β requires that $t_i \in \beta_i(t_i)$ for all $t_i \in T_i$ and $i \in I$. Thus, there is a unique set-valued deception that is single-valued, that is, β such that $\beta_i(t_i) = \{t_i\}$ for all $t_i \in T_i$ and $i \in I$. Hence, any single-valued deception β^s such that $\beta_i^s(t_i) \neq t_i$ for some $t_i \in T_i$ and $i \in I$ cannot be expressed as a special case of our set-valued deceptions.

Next, we recall another necessary condition for full implementation in Bayesian equilibrium.

Definition 9. An SCF f satisfies *Bayesian monotonicity (BM)* if, for every single-valued deception β^s that is unacceptable for f, there exist $i \in I$, $t_i \in T_i$, and $y : T_{-i} \to \Delta(A)$ such that

$$U_i(y \circ \beta_{-i}^s | t_i) > U_i(f \circ \beta^s | t_i),$$

whereas for all $\tilde{t}_i \in T_i$,

$$U_i(f | \tilde{t}_i) \geq U_i(y | \tilde{t}_i).$$

By undermining an unacceptable deception, as with weak IRM or IRM, type t_i can be used as an ally to a designer who wishes to implement f, this time in Bayesian equilibrium. However, because equilibrium (as opposed to rationalizability) is the solution concept used, the deceptions considered in BM are single-valued and the requirements on beliefs over the preference reversal are significantly reduced. As a consequence, Bergemann and Morris [7] are able to show that IRM implies BM. However, as we show in Examples 2 and 6, BM is not necessarily weaker than weak IRM.

6. Sufficiency for Implementation of an SCF in Interim Rationalizable Strategies

In this section, we show that weak IRM is sufficient for implementation in interim rationalizable strategies under a mild additional assumption: weak no-worst-rule (weak NWR).

For each $i \in I$ and $t_i \in T_i$, define

$$Y_i^w[t_i,f] \equiv \{y : T_{-i} \to \Delta(A) : U_i(f|t_i) \geq U_i(y|t_i)\}. \tag{1}$$

Thus, $Y_i^w[t_i,f]$ is the collection of all mappings $y : T_{-i} \to \Delta(A)$ such that y is weakly worse than f for individual i of type t_i. Notice that $Y_i[t_i,f]$ is a subset of $Y_i^w[t_i,f]$.)

Definition 10. The SCF f satisfies the *weak no-worst-rule* condition (weak NWR) if, for all $i \in I$, $t_i \in T_i$, and $\phi_i \in \Delta(T_{-i} \times T_{-i})$, there exist $y, y' \in Y_i^w[t_i,f]$ such that

$$\sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i}) u_i\big(y(t'_{-i}),(t_i,t_{-i})\big) \neq \sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i}) u_i\big(y'(t'_{-i}),(t_i,t_{-i})\big).$$

Remark. The weak NWR condition implies that the strictly lower contour set of f is nonempty for all types. Kunimoto [22] also defines a "no-worst-rule" condition that is stronger than our definition. Kunimoto [22] requires the existence of mappings y and y' in the set $\cap_{\tilde{t}_i \in T_i} Y_i^w[\tilde{t}_i,f]$, whereas we only require the existence of y and y' in the set $Y_i^w[t_i,f]$.

In the sufficiency result, we focus on a countable subset of $Y_i^w[t_i,f]$, as defined next. We denote by $\Delta^*(A)$ a countable dense subset of $\Delta(A)$. For each $i \in I$ and $t_i \in T_i$, define

$$Y_i^*[t_i,f] \equiv \left\{ y : T_{-i} \to \Delta(A) : \begin{array}{ll} \text{(i)} & y(t_{-i}) \in \Delta^*(A) \bigcup\limits_{t'_i \in T_i} \{f(t'_i,t_{-i})\},\ \forall t_{-i} \in T_{-i}, \text{ and} \\ \text{(ii)} & U_i(f|t_i) \geq U_i(y|t_i). \end{array} \right\}$$

Note that $Y_i^*[t_i,f] \subseteq Y_i^w[t_i,f]$. Because T_{-i} is finite and $\Delta^*(A) \cup_{t'_i \in T_i} \{f(t'_i,t_{-i})\}$ is countable, $Y_i^*[t_i,f]$ is also countable. Thus, we denote $Y_i^*[t_i,f]$ by $\{y_i^0[t_i,f], y_i^1[t_i,f], \ldots, y_i^k[t_i,f], \ldots\}$. For each $i \in I$ and $t_i \in T_i$, we then define $y_i^{t_i,f}$ such that

$$y_i^{t_i,f}(t_{-i}) = (1-\delta)\sum_{k=0}^{\infty} \delta^k y_i^k[t_i,f](t_{-i}),\ \forall t_{-i},$$

where $\delta \in (0,1)$.

Similarly, because A is countable, we denote it by $\{a_0, a_1, \ldots, a_k, \ldots\}$. Then, we define

$$\overline{\alpha} = (1-\eta)\sum_{k=0}^{\infty} \eta^k a_k,$$

where $\eta \in (0,1)$.

The following lemma notes two important consequences of weak NWR.

Lemma 3. *If an SCF f satisfies weak NWR, then the following statements are true:*

(a) *For all $i \in I$, $t_i \in T_i$, and $\phi_i \in \Delta(T_{-i} \times T_{-i})$, there exists $y \in Y_i^*[t_i,f]$ such that*

$$\sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i}) u_i\big(y(t'_{-i}),(t_i,t_{-i})\big) > \sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i}) u_i(y_i^{t_i,f}(t'_{-i}),(t_i,t_{-i})).$$

(b) *For all $i \in I$, $t_i \in T_i$, and $z_i^1 \in \Delta(T_{-i})$, there exists $a \in A$ such that*

$$\sum_{t_{-i}} z_i^1(t_{-i}) u_i(a,(t_i,t_{-i})) > \sum_{t_{-i}} z_i^1(t_{-i}) u_i(\overline{\alpha},(t_i,t_{-i})).$$

We now state and prove our sufficiency result for implementation in interim rationalizable strategies.

Theorem 2. *For any SCF f, if there exists an SCF $\hat{f} \approx f$ such that $\hat{f}$ satisfies weak IRM and weak NWR, then the SCF f is implementable in interim rationalizable strategies.*

The example shows that weak NWR is *not* necessary for interim rationalizable implementation:

Example 1. Let there be two agents, that is, $I = \{1, 2\}$. Agent 1 is uninformed of the state, so that he has one possible type t_1. Agent 2 is informed of the state, so that he has two possible types, t_2 and t'_2. There are three alternatives, that is, $A = \{a_1, a_2, a_3\}$. Agent 1 has state-independent preferences such that a_1 is strictly better than a_2, which is strictly better than a_3. Agent 2's his preferences in state (t_1, t_2) are a_3 better than a_2 better than a_1, and those in state (t_1, t'_2), a_3 better than a_1 better than a_2. Let f be the constant SCF, prescribing a_3 in both states.

Consider a mechanism in which agent 2 dictates an alternative, and it is implemented. Obviously, f is implementable in interim rationalizable strategies, which, by our Theorem 1, implies that f satisfies weak IRM. However, f violates weak NWR.

Hence, although weak NWR could be dispensed with in some environments, and although we consider it a very mild condition, we currently do not know whether weak NWR is indispensable in our Theorem 2.

7. Significance of the Results

We have shown that weak IRM is necessary and almost sufficient for implementation in interim rationalizable strategies (Theorems 1 and 2). In this section, we demonstrate the significance of these results in economically meaningful classes of environments.

Here is the roadmap for this section. In Section 7.1, we focus on private values environments and prove that many socially desirable SCFs are interim rationalizable implementable. First, acknowledging the difficulty of checking weak IRM directly, we propose an easy-to-check condition, which we term COND-1. Together with SIRBIC, COND-1 implies weak IRM. Second, we present a political-economy example with a class of nonconstant SCFs that always violate IRM. Then, using the workable COND-1, we show that a subset of those SCFs are implementable in interim rationalizable strategies. Furthermore, by focusing on this subset of SCFs, we show that there are some circumstances in which these SCFs satisfy BM, whereas in other circumstances, these SCFs violate it and therefore are not Bayesian implementable. Third, when we focus on interior SCFs, that is, those assigning positive probability to every pure outcome, we show that two mild conditions, closely related to those found in the literature, are sufficient for interim rationalizable implementation. Applying this result to a model of bilateral trading, we show that any SCF that satisfies BIC is "virtually" implementable (i.e., implementable with arbitrarily high probability) in interim rationalizable strategies. Fourth, we consider strategy-proof SCFs. COND-1, together with strategy-proofness, weak nonbossiness, and weak NWR, is sufficient for interim rationalizable implementation. We apply this result to two well-known classes of economic environments (pure exchange economies and a market with indivisible objects).

Finally, in Section 7.2, we also show that there exists an SCF that violates BM but is implementable in interim rationalizable strategies in an interdependent values environment. We also argue that this example is robust to the perturbation of the environment in an open set of utilities or beliefs. Thus, our findings in the previous subsection are extended to interdependent values environments.

7.1. Private Values Environments

We speak of a *private values* environment whenever each agent's preferences are independent of the other agents' types. Formally, the von Neumann-Morgenstern utility function of each agent i of type t_i is expressed as follows: for any $t_{-i}, t'_{-i} \in T_{-i}$ and $\ell \in \Delta(A)$,

$$u_i(\ell, (t_i, t_{-i})) = u_i(\ell, (t_i, t'_{-i})).$$

For simplicity, we then write the utility function as $u_i(\cdot, t_i)$.

7.1.1. Easy-to-Check Sufficient Condition. For interim rationalizable implementation, the key condition is weak IRM, which may often be tedious to check directly. We therefore propose a simple condition, which together with SIRBIC implies weak IRM.

Definition 11. The SCF f satisfies *COND-1* (*condition 1*) if, for any $i \in I$ and $t_i, t'_i \in T_i$, if $t'_i \nsim^f_i t_i$, then there exists $y \in Y^w_i[t'_i, f]$ such that, for any $t_{-i} \in T_{-i}$,

$$u_i(y(t_{-i}), t_i) > u_i(f(t'_i, t_{-i}), t_i).$$

COND-1 requires that whenever $t_i \nsim^f_i t'_i$, there be a preference reversal with respect to the SCF f between two types t_i and t'_i such that the strict inequality holds for type t_i in terms of "ex post" preferences, whereas the weak inequality holds for type t'_i in terms of "interim" preferences. This is weaker than requiring ex post preference reversals between the two types.

Proposition 1. *In a private values environment, if the SCF f satisfies SIRBIC and COND-1, it satisfies weak IRM.*

Remark. This result shows that, in a private values environment, weak IRM can roughly be decomposed into the "incentive compatibility" part (i.e., SIRBIC) and the "preference reversal" part (i.e., COND-1). These two distinct parts can also be found in the necessary conditions for Bayesian implementation: BIC and BM correspond to the incentive compatibility part and the preference reversal part, respectively.

We then obtain the following as a corollary of the previous proposition and Theorem 2.

Corollary 2. *In a private values environment, for any SCF f, if there exists an SCF $\hat{f} \approx f$ such that $\hat{f}$ satisfies SIRBIC, COND-1, and weak NWR, then f is implementable in interim rationalizable strategies.*

We now present a political-economy example with a class of nonconstant SCFs that always violate IRM. Then, we show that a subset of those SCFs, parameterized by $\epsilon \in (0,1)$, satisfy the sufficient conditions in Corollary 2, and hence, every SCF in that subset is implementable in interim rationalizable strategies. This disproves the claim made in Oury and Tercieux ([33], footnote 4) that IRM is necessary for interim rationalizable implementation. In the rest of the example, we focus on the subset of SCFs parametrized by ϵ. First, Claim 3 below exhibits circumstances in which every SCF in the subset satisfies BM. Second, Claim 4 below provides other circumstances in which every SCF in the subset violates BM, and thus those SCFs are not Bayesian implementable.[17]

Example 2 (Gaps Between IRM, BM, and Weak IRM). There are two players, that is, $I = \{1,2\}$. There is a set of five alternatives $A = \{a,b,c,d,e\}$ on the real line, which can be interpreted as political positions, with the following ordering: $a < b < c < d < e$. Each player $i \in I$ has three types: $T_i = \{t_i, t_i', t_i''\}$. Each type's preferences on A are single-peaked. Types t_i and t_i' of player i are extremists, whereas type t_i'' is a moderate. Specifically, type t_i is a left extremist, whose preferences peak at a. In contrast, type t_i' is a right extremist, whose preferences peak at e. Type t_i'' of player i is a moderate but left-of-center, whose preferences peak at b.

For each player $i \in I$, Table 1 lists the payoffs of the extremist and moderate types:

An extremist believes that the other player is also an extremist, with equal probability of being either left or right extremist. Thus, the belief of type t_i of player i is $\pi_i(t_i)[t_{-i}] = \pi_i(t_i)[t'_{-i}] = 0.5$. Likewise, the belief of type t_i' of player i is $\pi_i(t_i')[t_{-i}] = \pi_i(t_i')[t'_{-i}] = 0.5$. In contrast, the moderate type t_i'' of player i believes that the other player is more likely to be a moderate than an extremist. That is, $\pi_i(t_i'')[t''_{-i}] \geq 0.5$.

We envision a planner with somewhat moderate goals. Specifically, throughout the example, we focus on the class of SCFs F that assign the center alternative c whenever both players are extremists, that is, if $f \in F$, then $f(t_1,t_2) = f(t_1,t_2') = f(t_1',t_2) = f(t_1',t_2') = c$. We now show that any nonconstant SCF in F violates IRM.

Claim 1. Let $f \in F$. If the SCF f is not constant, it violates IRM.

Proof. Let f be the SCF such that $f \in F$ and it is not constant. Consider the deception β such that $\beta_i(t_i) = \{t_i\}$, $\beta_i(t_i') = \{t_i'\}$ and $\beta_i(t_i'') = \{t_i, t_i', t_i''\}$ for all $i \in I$. Because $f \in F$ and f is not constant, it follows that there exist $i \in I$ and $\hat{t}_{-i} \in T_{-i}$ such that $f(t_i'', \hat{t}_{-i}) \neq c$. However, regardless of the actual value of $\hat{t}_{-i}$, there always exists $\tilde{t} \in \beta_i(t_i'') \times \beta_{-i}(\hat{t}_{-i})$ such that $f(\tilde{t}) = c$. Hence, β is unacceptable for f.

Suppose, by way of contradiction, that the SCF f satisfies IRM. Then there exist $i \in I$, $\hat{t}_i \in T_i$, and $\hat{t}_i' \in \beta_i(\hat{t}_i)$ satisfying $\hat{t}_i' \nsim_i^f \hat{t}_i$ such that for all $\phi_i \in \Delta(T_{-i} \times T_{-i})$ satisfying $\phi_i(\hat{t}_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(\hat{t}_{-i})$ and $\pi_i(\hat{t}_i)[\hat{t}_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(\hat{t}_{-i}, \tilde{t}_{-i})$ for all $\hat{t}_{-i} \in T_{-i}$, there exists $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such that

$$\sum_{\hat{t}_{-i}, \tilde{t}_{-i}} \phi_i(\hat{t}_{-i}, \tilde{t}_{-i}) u_i(y(\tilde{t}_{-i}), \hat{t}_i) > \sum_{\hat{t}_{-i}, \tilde{t}_{-i}} \phi_i(\hat{t}_{-i}, \tilde{t}_{-i}) u_i(f(\hat{t}_i', \tilde{t}_{-i}), \hat{t}_i).$$

Because $\beta_i(t_i) = \{t_i\}$ and $\beta_i(t_i') = \{t_i'\}$, it must be that $\hat{t}_i = t_i''$ and $\hat{t}_i' \in \{t_i, t_i'\}$. Now consider the belief $\phi_i \in \Delta(T_{-i} \times T_{-i})$ such that $\phi_i(t_{-i}, t_{-i}) = \pi_i(t_i'')[t_{-i}]$, $\phi_i(t'_{-i}, t'_{-i}) = \pi_i(t_i'')[t'_{-i}]$, and $\phi_i(t''_{-i}, t_{-i}) + \phi_i(t''_{-i}, t'_{-i}) = \pi_i(t_i'')[t''_{-i}]$ (i.e., $\phi_i(t''_{-i}, t''_{-i}) = 0$) such that

$$\phi_i(t''_{-i}, t_{-i}) + \phi_i(t_{-i}, t_{-i}) = \phi_i(t''_{-i}, t'_{-i}) + \phi_i(t'_{-i}, t'_{-i}) = 0.5.$$

Table 1. The payoffs of the extremist and moderate types in Example 2.

	a	b	c	d	e
t_i	1	3/4	1/2	1/16	0
t_i'	0	2/5	1/2	3/4	1
t_i''	1/16	1	3/4	1/2	0

Such a ϕ_i exists because, by assumption, $\pi_i(t_i'')[t_{-i}''] \geq 0.5$. Furthermore, ϕ_i satisfies the requirement that $\phi_i(\hat{t}_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(\hat{t}_{-i})$ and $\pi_i(t_i'')[\hat{t}_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(\hat{t}_{-i}, \tilde{t}_{-i})$ for all $\hat{t}_{-i} \in T_{-i}$. Thus, there must exist a $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such that

$$\sum_{\hat{t}_{-i}, \tilde{t}_{-i}} \phi_i(\hat{t}_{-i}, \tilde{t}_{-i}) u_i(y(\tilde{t}_{-i}), t_i'') > \sum_{\hat{t}_{-i}, \tilde{t}_{-i}} \phi_i(\hat{t}_{-i}, \tilde{t}_{-i}) u_i(f(\tilde{t}_i', \tilde{t}_{-i}), t_i'')$$
$$\Leftrightarrow 0.5 u_i(y(t_{-i}), t_i'') + 0.5 u_i(y(t_{-i}'), t_i'') > u_i(c, t_i'').$$

As $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$, we must satisfy two other inequalities: $U_i(f|t_i) \geq U_i(y|t_i)$ and $U_i(f|t_i') \geq U_i(y|t_i')$. The three inequalities together boil down to

$$0.5 u_i(y(t_{-i}), t_i'') + 0.5 u_i(y(t_{-i}'), t_i'') > \frac{3}{4};$$
$$\frac{1}{2} \geq 0.5 u_i(y(t_{-i}), t_i) + 0.5 u_i(y(t_{-i}'), t_i); \text{ and}$$
$$\frac{1}{2} \geq 0.5 u_i(y(t_{-i}), t_i') + 0.5 u_i(y(t_{-i}'), t_i').$$

Let $\ell_1 = y(t_{-i})$ and $\ell_2 = y(t_{-i}')$ be the lotteries we consider. Then, the previous inequalities translate into, respectively,

$$\frac{1}{16}(\ell_1[a] + \ell_2[a]) + (\ell_1[b] + \ell_2[b]) + \frac{3}{4}(\ell_1[c] + \ell_2[c]) + \frac{1}{2}(\ell_1[d] + \ell_2[d]) > \frac{3}{2};$$
$$1 \geq (\ell_1[a] + \ell_2[a]) + \frac{3}{4}(\ell_1[b] + \ell_2[b]) + \frac{1}{2}(\ell_1[c] + \ell_2[c]) + \frac{1}{16}(\ell_1[d] + \ell_2[d]); \text{ and}$$
$$1 \geq \frac{2}{5}(\ell_1[b] + \ell_2[b]) + \frac{1}{2}(\ell_1[c] + \ell_2[c]) + \frac{3}{4}(\ell_1[d] + \ell_2[d]) + (\ell_1[e] + \ell_2[e]),$$

where $\ell_k[x]$ denotes the probability that alternative $x \in A$ is chosen under the lottery ℓ_k. However, it is impossible to simultaneously satisfy these three inequalities, a contradiction. □

Next, let the lottery $\ell^* = 2a/5 + 3d/5$ (Table 2). Fix any $\epsilon \in (0,1)$, and consider the following SCF f^ϵ:

Notice that $f^\epsilon \in F$ for all $\epsilon \in (0,1)$. We now show that f^ϵ satisfies the sufficient condition in Corollary 2, and hence it is implementable in interim rationalizable strategies.

Claim 2. For all $\epsilon \in (0,1)$, the SCF f^ϵ is implementable in interim rationalizable strategies.

Proof. It is straightforward to check that f^ϵ satisfies SIRBIC. (All types strictly prefer c to ℓ^*. Moreover, the incentive constraints are strict for type t_i'' because, by assumption, $\pi_i(t_i'')[t_{-i}''] \geq 0.5$.) It can also be easily checked that f^ϵ satisfies weak NWR because for all $\hat{t} \in T$ and $i \in I$, there exists an alternative that is strictly worse than $f^\epsilon(\hat{t})$ for type $\hat{t}_i$.

Finally, we claim that f^ϵ satisfies COND-1. Fix $i \in I$. Let $y : T_{-i} \rightarrow \Delta(A)$ be such that $y(\hat{t}_{-i}) = a$ for all $\hat{t}_{-i} \in T_{-i}$. By construction of y, we confirm that $y \in Y_i^w[t_i'', f^\epsilon]$ and $u_i(y(\hat{t}_{-i}), t_i) > u_i(f^\epsilon(t_i'', \hat{t}_{-i}), t_i)$ for all $\hat{t}_{-i} \in T_{-i}$. Hence, COND-1 holds for the case that $t_i'' \not\sim_i^{f^\epsilon} t_i$. By a similar argument (e.g., letting $y(\hat{t}_{-i}) = e$ for all $\hat{t}_{-i} \in T_{-i}$), COND-1 also holds for the case that $t_i'' \not\sim_i^{f^\epsilon} t_i'$.

Next, let $y : T_{-i} \rightarrow \Delta(A)$ be such that $y(\hat{t}_{-i}) = 7b/11 + 4d/11$ for all $\hat{t}_{-i} \in T_{-i}$. By construction of y, we confirm that $y \in Y_i^w[t_i, f^\epsilon]$ and $u_i(y(\hat{t}_{-i}), t_i'') > u_i(f^\epsilon(t_i, \hat{t}_{-i}), t_i'')$ for all $\hat{t}_{-i} \in T_{-i}$. Hence, COND-1 holds for the case that $t_i \not\sim_i^{f^\epsilon} t_i''$.

Last, let $y : T_{-i} \rightarrow \Delta(A)$ such that $y(\hat{t}_{-i}) = b$ for all $\hat{t}_{-i} \in T_{-i}$. By construction of y, we confirm that $y \in Y_i^w[t_i', f^\epsilon]$ and $u_i(y(\hat{t}_{-i}), t_i'') > u_i(f^\epsilon(t_i', \hat{t}_{-i}), t_i'')$ for all $\hat{t}_{-i} \in T_{-i}$. Hence, COND-1 holds for the case that $t_i' \not\sim_i^{f^\epsilon} t_i''$. This completes the proof that f^ϵ satisfies COND-1.

It follows from Corollary 2 that f^ϵ is implementable in interim rationalizable strategies. □

Table 2. The description of the SCF f^ϵ in Example 2.

f^ϵ	t_2	t_2'	t_2''
t_1	c	c	$(1-\epsilon)c + \epsilon\ell^*$
t_1'	c	c	$(1-\epsilon)c + \epsilon\ell^*$
t_1''	$(1-\epsilon)c + \epsilon\ell^*$	$(1-\epsilon)c + \epsilon\ell^*$	b

The previous two claims demonstrate that IRM can be strictly stronger than weak IRM and the former condition is not necessary for interim rationalizable implementation. The next claim establishes that IRM can be strictly stronger than BM. Because $f^{\epsilon} \in F$ and it is not constant, it follows from Claim 1 that f^{ϵ} does not satisfy IRM for all $\epsilon \in (0, 1)$. However, now we show that if $\pi_i(t''_i)[t''_{-i}] = 1$ for all $i \in I$, then f^{ϵ} satisfies BM.

Claim 3. *Suppose $\pi_i(t''_i)[t''_{-i}] = 1$ for all $i \in I$. The SCF f^{ϵ} satisfies BM for all $\epsilon \in (0, 1)$.*

Proof. Consider any single-valued deception β^s such that β^s is unacceptable for f^{ϵ}. As β^s is unacceptable, there must exist $j \in I$ such that at least one of the following is true: $\beta^s_j(t''_j) \neq t''_j$, $\beta^s_j(t_j) = t''_j$ or $\beta^s_j(t'_j) = t''_j$.

First, suppose $\beta^s_j(t''_j) \neq t''_j$ for some $j \in I$. There are three possible subcases to consider here:

Subcase 1. Suppose $\beta^s_j(t''_j) \neq t''_j$ and $\beta^s_{-j}(t''_{-j}) = t''_{-j}$. Then $f^{\epsilon}(\beta^s_j(t''_j), \beta^s_{-j}(t''_{-j})) = (1-\epsilon)c + \epsilon\ell^*$. Let $y : T_{-j} \longrightarrow \Delta(A)$ be such that $y(t_{-j}) = y(t'_{-j}) = (1-\epsilon)c + \epsilon\ell^*$ and $y(t''_{-j}) = b$. Then

$$U_j(y \circ \beta^s_{-j} | t''_j) > U_j(f^{\epsilon} \circ \beta^s | t''_j),$$

whereas $U_j(f^{\epsilon} | \hat{t}_j) \geq U_j(y | \hat{t}_j)$ for all $\hat{t}_j \in T_j$.

Subcase 2. Suppose $\beta^s_j(t''_j) \neq t''_j$ and $\beta^s_{-j}(t''_{-j}) = t_{-j}$. Then $f^{\epsilon}(\beta^s_j(t''_j), \beta^s_{-j}(t''_{-j})) = c$. Let $y : T_{-j} \longrightarrow \Delta(A)$ be such that $y(t_{-j}) = y(t''_{-j}) = b$ and $y(t'_{-j}) = a/5 + 4d/5$. Then

$$U_j(y \circ \beta^s_{-j} | t''_j) > U_j(f^{\epsilon} \circ \beta^s | t''_j),$$

whereas $U_j(f^{\epsilon} | \hat{t}_j) \geq U_j(y | \hat{t}_j)$ for all $\hat{t}_j \in T_j$.

Subcase 3. Suppose $\beta^s_j(t''_j) \neq t''_j$ and $\beta^s_{-j}(t''_{-j}) = t'_{-j}$. Then $f^{\epsilon}(\beta^s_j(t''_j), \beta^s_{-j}(t''_{-j})) = c$. Let $y : T_{-j} \longrightarrow \Delta(A)$ be such that $y(t'_{-j}) = y(t''_{-j}) = b$ and $y(t_{-j}) = a/5 + 4d/5$. Then

$$U_j(y \circ \beta^s_{-j} | t''_j) > U_j(f^{\epsilon} \circ \beta^s | t''_j),$$

whereas $U_j(f^{\epsilon} | \hat{t}_j) \geq U_j(y | \hat{t}_j)$ for all $\hat{t}_j \in T_j$.

Second, suppose $\beta^s_i(t''_i) = t''_i$ for all $i \in I$ but $\beta^s_j(t_j) = t''_j$ for some $j \in I$. There are two possible subcases to consider here:

Subcase 1. Suppose both $\beta^s_{-j}(t_{-j})$ and $\beta^s_{-j}(t'_{-j})$ are in $\{t_{-j}, t'_{-j}\}$. Then $f^{\epsilon}(\beta^s_j(t_j), \beta^s_{-j}(t_{-j})) = f^{\epsilon}(\beta^s_j(t_j), \beta^s_{-j}(t'_{-j})) = (1-\epsilon)c + \epsilon\ell^*$. Let $y : T_{-j} \longrightarrow \Delta(A)$ be such that $y(\hat{t}_{-j}) = c$ for all $\hat{t}_{-j} \in T_{-j}$. Then

$$U_j(y \circ \beta^s_{-j} | t_j) > U_j(f^{\epsilon} \circ \beta^s | t_j),$$

whereas $U_j(f^{\epsilon} | \hat{t}_j) \geq U_j(y | \hat{t}_j)$ for all $\hat{t}_j \in T_j$.

Subcase 2. Suppose there exists $\hat{t}_{-j} \in \{t_{-j}, t'_{-j}\}$ such that $\beta^s_{-j}(\hat{t}_{-j}) = t''_{-j}$. Without loss of generality, suppose $\hat{t}_{-j} = t_{-j}$. Then $f^{\epsilon}(\beta^s_j(t_j), \beta^s_{-j}(t_{-j})) = b$ and $f^{\epsilon}(\beta^s_j(t_j), \beta^s_{-j}(t'_{-j})) \in \{b, (1-\epsilon)c + \epsilon\ell^*\}$. Let $y : T_{-j} \longrightarrow \Delta(A)$ be such that $y(t_{-j}) = y(t'_{-j}) = c$ and $y(t''_{-j}) = a$. Then

$$U_j(y \circ \beta^s_{-j} | t_j) > U_j(f^{\epsilon} \circ \beta^s | t_j),$$

whereas $U_j(f^{\epsilon} | \hat{t}_j) \geq U_j(y | \hat{t}_j)$ for all $\hat{t}_j \in T_j$.

Finally, suppose $\beta^s_i(t''_i) = t''_i$ for all $i \in I$ but $\beta^s_j(t'_j) = t''_j$ for some $j \in I$. There are two possible subcases to consider here:

Subcase 1. Suppose both $\beta^s_{-j}(t_{-j})$ and $\beta^s_{-j}(t'_{-j})$ are in $\{t_{-j}, t'_{-j}\}$. Then $f^{\epsilon}(\beta^s_j(t'_j), \beta^s_{-j}(t_{-j})) = f^{\epsilon}(\beta^s_j(t'_j), \beta^s_{-j}(t'_{-j})) = (1-\epsilon)c + \epsilon\ell^*$. Let $y : T_{-j} \longrightarrow \Delta(A)$ be such that $y(\hat{t}_{-j}) = c$ for all $\hat{t}_{-j} \in T_{-j}$. Then

$$U_j(y \circ \beta^s_{-j} | t'_j) > U_j(f^{\epsilon} \circ \beta^s | t'_j),$$

whereas $U_j(f^{\epsilon} | \hat{t}_j) \geq U_j(y | \hat{t}_j)$ for all $\hat{t}_j \in T_j$.

Subcase 2. Suppose there exists $\hat{t}_{-j} \in \{t_{-j}, t'_{-j}\}$ such that $\beta^s_{-j}(\hat{t}_{-j}) = t''_{-j}$. Without loss of generality, suppose $\hat{t}_{-j} = t_{-j}$. Then $f^{\epsilon}(\beta^s_j(t'_j), \beta^s_{-j}(t_{-j})) = b$ and $f^{\epsilon}(\beta^s_j(t'_j), \beta^s_{-j}(t'_{-j})) \in \{b, (1-\epsilon)c + \epsilon\ell^*\}$. Let $y : T_{-j} \longrightarrow \Delta(A)$ be such that $y(t_{-j}) = y(t'_{-j}) = c$ and $y(t''_{-j}) = e$. Then

$$U_j(y \circ \beta^s_{-j} | t'_j) > U_j(f^{\epsilon} \circ \beta^s | t'_j),$$

whereas $U_j(f^{\epsilon} | \hat{t}_j) \geq U_j(y | \hat{t}_j)$ for all $\hat{t}_j \in T_j$.

It follows from the previous arguments that f^{ϵ} satisfies BM. □

Thus, as a consequence of the previous claims, when $\pi_i(t_i'')[t_{-i}''] = 1$ for all $i \in I$, the SCF f^ϵ satisfies both BM and weak IRM but not IRM.

In the final claim, we show that, when $\pi_i(t_i'')[t_{-i}] = \pi_i(t_i'')[t_{-i}''] = 1/2$ for all $i \in I$, the SCF f^ϵ violates BM. However, as f^ϵ satisfies weak IRM under those conditions, the claim demonstrates that weak IRM can be strictly weaker than BM too and that there are SCFs that are implementable in interim rationalizable strategies but not in Bayesian equilibrium.

Claim 4. Suppose $\pi_i(t_i'')[t_{-i}'] = \pi_i(t_i'')[t_{-i}''] = 1/2$ for all $i \in I$. The SCF f^ϵ violates BM for all $\epsilon \in (0,1)$.

Proof. Consider the single-valued deception β^s such that $\beta_i^s(t_i) = \beta_i^s(t_i'') = t_i$ and $\beta_i^s(t_i') = t_i'$ for all $i \in I$. Then $f \circ \beta^s(\hat{t}) = c$ for all $\hat{t} \in T$. Hence, β^s is unacceptable for f^ϵ.

Suppose, by way of contradiction, that the SCF f^ϵ satisfies BM. Then there exist $i \in I$, $\tilde{t}_i \in T_i$, and $y : T_{-i} \to \Delta(A)$ such that

$$U_i(y \circ \beta_{-i}^s | \tilde{t}_i) > U_i(f^\epsilon \circ \beta^s | \tilde{t}_i),$$

whereas $U_i(f^\epsilon | \hat{t}_i) \geq U_i(y | \hat{t}_i)$ for all $\hat{t}_i \in T_i$.

It cannot be the case that $\tilde{t}_i$ is either equal to t_i or t_i'. For instance, suppose $\tilde{t}_i = t_i$. Then $U_i(y \circ \beta_{-i}^s | t_i) = U_i(y | t_i)$ and $U_i(f^\epsilon \circ \beta^s | t_i) = U_i(f^\epsilon | t_i)$. Hence, we cannot obtain the preference reversal required by BM. The same argument applies to the case when $\tilde{t}_i = t_i'$.

Therefore, suppose $\tilde{t}_i = t_i''$. Then $U_i(y \circ \beta_{-i}^s | t_i'') > U_i(f^\epsilon \circ \beta^s | t_i'')$, $U_i(f^\epsilon | t_i) \geq U_i(y | t_i)$, and $U_i(f^\epsilon | t_i') \geq U_i(y | t_i')$ are equivalent to, respectively,

$$0.5u_i(y(t_{-i}), t_i'') + 0.5u_i(y(t_{-i}'), t_i'') > \frac{3}{4};$$

$$\frac{1}{2} \geq 0.5u_i(y(t_{-i}), t_i) + 0.5u_i(y(t_{-i}'), t_i); \text{ and}$$

$$\frac{1}{2} \geq 0.5u_i(y(t_{-i}), t_i') + 0.5u_i(y(t_{-i}'), t_i').$$

However, as mentioned in the proof of Claim 1, it is impossible to simultaneously satisfy the three inequalities, a contradiction. □

We compare the results in this class of examples with Oury and Tercieux [33], which proves that IRM is necessary for strict continuous (partial) implementation in Bayesian equilibrium. Our examples show that, although the SCF f^ϵ is interim rationalizable implementable, it cannot be strictly continuously implemented in Bayesian equilibrium because it fails IRM. Thus, any mechanism that implements f^ϵ in interim rationalizable strategies must exhibit either of the following two types of failures of strict continuous implementation: either nonexistence of strict Bayesian equilibria, or the lack of a "continuous" extension of strict Bayesian equilibria to nearby environments where a continuous extension is defined as a continuous mapping from the universal type space endowed with the product topology to the space of lotteries over outcomes endowed with the standard Euclidean metric. That is, even if there exist strict Bayesian equilibria in the implementing mechanism, any strict Bayesian equilibrium in the original incomplete information environment *cannot* be extended to a nearby incomplete information environment as a Bayesian equilibrium in a "continuous" manner.

7.1.2. Interior SCFs. By the *interior of* $\Delta(A)$, we mean the set of all lotteries that assign a positive probability to all $a \in A$. We now show that for SCFs whose range is in the interior of $\Delta(A)$, both COND-1 and weak NWR are satisfied under mild conditions, that is, the environment satisfies "no-total-indifference" and the SCF is "responsive only when preferences differ."

The environment satisfies *no-total-indifference (NTI)* if, for all $i \in I$ and $t_i \in T_i$, there exist alternatives $a, a' \in A$ such that

$$u_i(a, t_i) \neq u_i(a', t_i).$$

For private-values environments, the previous definition is equivalent to the no-total-indifference condition introduced in Serrano and Vohra [40].

The SCF f is *responsive only when preferences differ* if, for any $i \in I$ and $t_i, t_i' \in T_i$, if $t_i \not\sim_i^f t_i'$, it follows that t_i and t_i' have different preferences on $\Delta(A)$; that is, $u_i(\cdot, t_i')$ is not a positive affine transformation of $u_i(\cdot, t_i)$. This condition is slightly weaker than the type diversity condition in Serrano and Vohra [40], as the latter requires *every* pair of distinct types to have different preferences on $\Delta(A)$.

Proposition 2. *In a private values environment satisfying NTI, if the SCF f is responsive only when preferences differ and f(t) is in the interior of $\Delta(A)$ for all $t \in T$, then f satisfies weak NWR and COND-1.*

We then obtain the following as a corollary of the above proposition and Corollary 2.

Corollary 3. *In a private values environment satisfying NTI, for any SCF f, if there exists an SCF $\hat{f} \approx f$ such that $\hat{f}$ satisfies SIRBIC, is responsive only when preferences differ, and $\hat{f}(t)$ is in the interior of $\Delta(A)$ for all $t \in T$, then f is implementable in interim rationalizable strategies.*

We now use this result to show that, in a model of bilateral trading, any SCF that satisfies BIC is "virtually" implementable in interim rationalizable strategies:

Example 3 (Bilateral Trading). Consider the model of bilateral trading from Myerson and Satterthwaite [29] except for the following modifications: We assume a discrete type space and monetary transfers that are rational numbers, allowing for correlated beliefs.

Formally, we consider a simple trading problem with two individuals, a seller, who has a single indivisible object to sell, and a buyer. Both attempt to agree on an exchange of the object for money. The buyer's type is equal to the value v whereas the seller's type is equal to the cost c. The buyer's value and the seller's cost are elements of a discrete grid on $[0,1]$. (This assumption makes the set of types finite.) The traders' beliefs are derived from a common prior π such that $\pi(v,c) > 0$ for all (v, c). Then $\pi_b(v)$ denotes the belief of type v of the buyer and $\pi_s(c)$ denotes the belief of type c of the seller.

Let one (zero) denote the event that the good is traded (not traded). We restrict the transfers to be rational numbers. Thus, $A = \{(x,y) : x \in \{0,1\} \text{ and } y \in \mathbb{Q}\}$. Note that A is countable.

The buyer's utility is $u_b((x,y),v) = vx - y$ and the seller's utility is $u_s((x,y),c) = y - cx$. Given these utility functions, the environment satisfies private values and NTI.

In this model, every SCF is responsive only when preferences differ. This is because the environment satisfies the slightly stronger condition of type diversity (Serrano and Vohra [40]), that is, $t_i \neq t_i'$ implies that t_i and t_i' have different preferences on $\Delta(A)$.

Consider any SCF f that satisfies BIC. We now argue that f is *virtually* implementable in interim rationalizable strategies, that is, for all $\epsilon > 0$, we can find an f^ϵ such that $\sup\{|f(v,c)[a] - f^\epsilon(v,c)[a]| : a \in A\} < \epsilon$, for all (v, c), and f^ϵ is implementable in interim rationalizable strategies. Here $f(v,c)[a]$ denotes the probability that outcome a is realized under $f(v, c)$.

The environment satisfies type diversity and NTI. Pick any $i \in I$ and $t_i, t_i' \in T_i$ such that $t_i \neq t_i'$. Then, as argued in the proof of Proposition 2, there exist lotteries $\ell', \ell'' \in \Delta(A)$ such that ℓ' and ℓ'' have finite supports and satisfy (A.10). Applying the lemma in Abreu and Matsushima [1], we get that for each $i \in I$, there exists $\ell_i : T_i \longrightarrow \Delta(A)$ such that for all $t_i \neq t_i'$,

$$u_i(\ell_i(t_i), t_i) > u_i(\ell_i(t_i'), t_i).$$

Pick any $\epsilon \in (0,1)$ and let $f^\epsilon(v,c) = (1-\epsilon)f(v,c) + \epsilon/3(\ell_b(v) + \ell_s(c) + \overline{\ell})$, for all (v,c), where $\overline{\ell}$ is any lottery that assigns a positive probability to all alternatives in A. Then f^ϵ satisfies SIRBIC, indeed, it satisfies strict BIC, is responsive only when preferences differ, and $f^\epsilon(v,c)$ is in the interior of $\Delta(A)$ for all (v, c). Hence, it follows from Corollary 3 that f^ϵ is implementable in interim rationalizable strategies. Thus, in the bilateral trading model, any SCF that satisfies BIC is virtually implementable in interim rationalizable strategies.

Corollary 1 of Serrano and Vohra [40] implies that, in the bilateral trading model, any SCF that satisfies BIC is virtually Bayesian implementable. Thus, we show that virtual implementation in interim rationalizable strategies is not less permissive in this setting.

7.1.3. Strategy-Proof SCFs. Strategy-proof SCFs have been extensively studied in the context of private values environments. The reader is referred to Barberà [3] for a survey on this literature. The SCF f is *strategy proof* if, for all $i \in I$, $t_i, t_i' \in T_i$, and $t_{-i} \in T_{-i}$,

$$u_i(f(t_i, t_{-i}), t_i) \geq u_i(f(t_i', t_{-i}), t_i).$$

The strategy proofness of f requires that telling their true type constitutes a dominant-strategy equilibrium in the direct mechanism associated with f. We now provide a sufficiency result for the implementation of strategy-proof SCFs in interim rationalizable strategies. We need one more definition before we present the result.

The SCF f satisfies *weak nonbossiness* if, for all $i \in I$ and $t_i, t_i' \in T_i$, if $t_i \not\sim_i^f t_i'$, then there exists $t_{-i}' \in T_{-i}$ such that $u_i(f(t_i, t_{-i}'), t_i) \neq u_i(f(t_i', t_{-i}'), t_i)$. See Thomson [44] for an extensive discussion on various versions of nonbossiness used in the literature.

Saijo et al. [38] and Mizukami and Wakayama [28] show that strategy-proofness and weak nonbossiness characterize SCFs that are dominant-strategy implementable by the associated direct mechanisms. However, dominant-strategy implementation does not rule out the possibility of "bad" equilibria or "bad" rationalizable strategies in the mechanism. The next result tells us that any such SCF that also satisfies COND-1 and weak NWR is implementable in interim rationalizable strategies in any environment where beliefs are strictly in the interior.

Proposition 3. *Suppose that in a private values environment, we have $\pi_i(t_i)[t_{-i}] > 0$ for all $t \in T$ and $i \in I$.*[18] *If the SCF f satisfies strategy-proofness, weak nonbossiness, COND-1, and weak NWR, then f is implementable in interim rationalizable strategies.*

Proof. Because f satisfies strategy proofness, it satisfies BIC. We first claim that f also satisfies SIRBIC. Pick any $i \in I$ and $t_i, t_i' \in T_i$ such that $t_i \not\approx_i^f t_i'$. Then weak nonbossiness implies that there exists $t_{-i}' \in T_{-i}$ such that $u_i(f(t_i, t_{-i}'), t_i) \neq u_i(f(t_i', t_{-i}'), t_i)$. By strategy proofness of f, it follows that $u_i(f(t_i, t_{-i}'), t_i) > u_i(f(t_i', t_{-i}'), t_i)$. Since $\pi_i(t_i)[t_{-i}'] > 0$ for any $t_{-i}' \in T_{-i}$, we conclude that f satisfies SIRBIC.

The result follows because f satisfies SIRBIC, COND-1, and weak NWR. □

We now present two applications of the foregoing result.

Example 4 (Pure-Exchange Economy). There are a finite number of goods, L. (We use L to denote both the set and number of goods.) Each individual $i \in I$ has an initial endowment e_i, where e_i^l denotes the endowment of good l. We assume that each e_i^l is a nonnegative rational number with $\sum_{l \in L} e_i^l > 0$, which means that individual i is endowed with a positive quantity of at least one good ℓ. The set of alternatives A is equal to $(a_1, \ldots, a_n) \in \mathbb{R}_+^{Ln}$ such that each a_i^l is a rational number and $\sum_{i \in I} a_i^l = \sum_{i \in I} e_i^l$ for all $l \in L$.

Each agent cares about only her own consumption. That is, for all $i \in I$, $t_i \in T_i$, and $a, \hat{a} \in A$, if $a_i = \hat{a}_i$, then $u_i(a, t_i) = u_i(\hat{a}, t_i)$. We assume that the agents' utility functions are strictly increasing in the quantity of each good. That is, for all $i \in I$, $t_i \in T_i$, and $a, \hat{a} \in A$, if $a_i^l \geq \hat{a}_i^l$ for all $\ell \in L$ but $a_i \neq \hat{a}_i$, then $u_i(a, t_i) > u_i(\hat{a}, t_i)$.

Here we focus only on individually rational SCFs. The SCF $f: T \to \Delta(A)$ is *individually rational* if $u_i(f(t), t_i) \geq u_i(e, t_i)$ for all $t \in T$, $t_i \in T_i$, and $i \in I$.

We confirm that any individually rational SCF f that satisfies strategy proofness and SIRBIC also satisfies COND-1 and weak NWR in our pure-exchange economy. Because f is individually rational, $\sum_{l \in L} e_i^l > 0$, for all $i \in I$, and utility functions are strictly increasing in the quantity of each good, it must be that for all $i \in I$ and $t \in T$, the allocation $f(t)$ assigns a positive quantity of some good to individual i. Let $\overline{a}(i)$ denote the allocation where individual i obtains the total endowment of the economy. COND-1 follows from strategy proofness and SIRBIC because $u_i(\overline{a}(i), t_i) > u_i(f(t'), t_i)$ for all $t' \in T$, $t_i \in T_i$, and $i \in I$.[19] Weak NWR follows because, for any $i, j \in I$ with $i \neq j$, we have that $u_i(f(t'), t_i) > u_i(\overline{a}(j), t_i)$ for all $t' \in T$, $t_i \in T_i$, and $i \in I$.

It follows from Proposition 3 that in a pure-exchange economy, if $\pi_i(t_i)[t_{-i}] > 0$, for all $t \in T$ and $i \in I$, then any SCF f that satisfies strategy-proofness, weak nonbossiness, and individual rationality is implementable in interim rationalizable strategies. (Recall from the proof of Proposition 3 that strategy proofness and weak nonbossiness imply SIRBIC when beliefs are in the interior.) For instance, the SCFs defined by the fixed-proportion anonymous trading are strategy proof, weakly nonbossy, and individually rational (Barberà and Jackson [4]). Hence, these SCFs are also implementable in interim rationalizable strategies whenever beliefs are strictly in the interior.

Example 5 (Market with Indivisible Goods). Each individual owns an indivisible good (e.g., a house). Let $e = (e_1, \ldots, e_n)$ denote the initial endowment, where e_i denotes the good owned by individual i. An *allocation* a is a permutation of the n goods over the n individuals. Thus, an allocation results in each individual receiving exactly one of the n indivisible goods. Let a_i denote the good assigned to individual i in the allocation a. We let $\mathcal{A}$ be the set of all allocations. This is an important class of economies, which has been investigated by many researchers since it was introduced by Shapley and Scarf [42].

Each agent cares about only the good assigned to her. That is, for all $i \in I$, $t_i \in T_i$, and allocations a and $\hat{a}$, if $a_i = \hat{a}_i$, then $u_i(a, t_i) = u_i(\hat{a}, t_i)$. We assume that the agents' have strict preferences over the goods. That is, for all $i \in I$, $t_i \in T_i$, and allocations a and $\hat{a}$, if $a_i \neq \hat{a}_i$, then $u_i(a, t_i) \neq u_i(\hat{a}, t_i)$.

For each $t \in T$, the core of this economy is nonempty, and it consists of a single allocation (Roth and Postlewaite [37]). Let f^c be the SCF such that $f^c(t)$ is the core allocation for all $t \in T$. Svensson [43] shows that f^c is strategy proof, weakly nonbossy, and individually rational (i.e., $u_i(f^c(t), t_i) \geq u_i(e, t_i)$ for all $t \in T$ and $i \in I$).

Now suppose the designer has the *ability to reward or punish any one of the agents* (e.g., through monetary transfers). Specifically, for any allocation $a \in \mathcal{A}$, let $\overline{a}(i)$ ($\underline{a}(i)$) denote the alternative in which every agent $j \in I$ receives the

Table 3. The payoff profiles in each type profile when alternative a is chosen in Example 6.

a	t_2	t_2'
t_1	4, 4	4, 0
t_1'	0, 0	4, 1
t_1''	1, 1	4, 0

same good a_j but agent i receives a reward (punishment) as well. The set of alternatives is defined as $A = \cup_{a \in \mathcal{A}, i \in I} \{a, \overline{a}(i), \underline{a}(i)\}$.

We now expand the agent's preferences to the set of alternatives A by assuming that each agent cares about the good assigned to herself and whether she is rewarded or punished. Thus, for all $a \in \mathcal{A}$, $t_i \in T_i$, and $i \in I$,

$$u_i(\overline{a}(i), t_i) > u_i(a, t_i) > u_i(\underline{a}(i), t_i)$$

$$\text{and } j \neq i \Rightarrow u_i(\overline{a}(j), t_i) = u_i(a, t_i) = u_i(\underline{a}(j), t_i).$$

Finally, the rewards and punishments are sufficiently small so that, for any $a, a' \in \mathcal{A}$, if $u_i(a, t_i) > u_i(a', t_i)$, then $u_i(a, t_i) > u_i(\underline{a}(i), t_i) > u_i(\overline{a}'(i), t_i) > u_i(a', t_i)$.

When the set of alternatives is expanded into A and the beliefs are strictly in the interior, the SCF f^c satisfies COND-1 and weak NWR. COND-1 can be shown using strategy proofness, SIRBIC (itself a consequence of strategy proofness and weak nonbossiness of f^c, as shown in the proof of Proposition 3), and the following: pick any $i \in I$ and $t_i \in T_i$. Let $a^*_{t_i}$ be type t_i's most-preferred allocation. Then $u_i(\overline{a}^*_{t_i}(i), t_i) > u_i(f^c(t'), t_i)$ for all $t' \in T$. (See the argument in Endnote 19.) Let $\underline{f}^c(t')(i)$ denote the alternative in which every agent $j \in I$ receives the same good as in $f^c(t')$ and agent i receives the punishment. Then, weak NWR follows because $u_i(f^c(t'), t_i) > u_i(\underline{f}^c(t')(i), t_i)$ for all $t' \in T$, $t_i \in T_i$, and $i \in I$.

It follows from Proposition 3 that in a market with indivisible goods, if $\pi_i(t_i)[t_{-i}] > 0$, for all $t \in T$ and $i \in I$, and the designer has the ability to reward or punish any one of the agents, then the SCF f^c that selects the core allocation in each state is implementable in interim rationalizable strategies.

7.2. Beyond Private Values Environments

In this section, we argue that the implications in Example 2 are extended from private values to interdependent values environments. We base our arguments on the following example, which is built on the example presented in Kunimoto and Saran [23].

Example 6. There are two players $i \in \{1, 2\}$. Player 1 has three types: $T_1 = \{t_1, t_1', t_1''\}$, and player 2 has two types: $T_2 = \{t_2, t_2'\}$. The beliefs of the players are as follows:

$$\pi_1(t_1)[t_2] = 0.99, \quad \pi_1(t_1')[t_2] = \pi_1(t_1'')[t_2] = 0$$

and

$$\pi_2(t_2)[t_1] = \pi_2(t_2)[t_1'] = \pi_2(t_2)[t_1''] = \frac{1}{3}, \quad \pi_2(t_2')[t_1'] = 1.$$

Notice that $T^* = T$ because for every state, there is a type of a player who puts positive probability on that state. Thus, we do not need to discuss equivalent SCFs in this example.

There are six pure alternatives: $A = \{a, b, c, d, z, z'\}$. Tables 3–8 list the payoffs of the two players:

The SCF f selects the alternative that maximizes the aggregate payoff in each state (Table 9).

Table 4. The payoff profiles in each type profile when alternative b is chosen in Example 6.

b	t_2	t_2'
t_1	0, 0	3, 3
t_1'	1, 1	2, 0
t_1''	0, 0	2, 1

Table 5. The payoff profiles in each type profile when alternative c is chosen in Example 6.

c	t_2	t_2'
t_1	0, 0	3, 1
t_1'	3, 3	3, 0
t_1''	3, 3	3, 0

Here we omit the derivation in all the claims we establish later.[20] We emphasize that the analysis based on this example is robust to small perturbations of the underlying beliefs and utility functions, as all the relevant conditions are expressed in terms of a finite number of strict inequalities.

Claim 5. The SCF f violates BM.

Because we know from Bergemann and Morris [7] that IRM implies BM, we also state the following.

Claim 6. The SCF f violates IRM.

In contrast, we have the following.

Claim 7. The SCF f satisfies weak IRM.

Claim 8. The SCF f satisfies weak NWR.

These two claims lead to the following.

Claim 9. The SCF f is implementable in interim rationalizable strategies by the canonical mechanism we used in Theorem 2.

Furthermore, there are no mixed Bayesian equilibria in that canonical mechanism.

Claim 10. There are no mixed Bayesian equilibria in the canonical mechanism implementing the SCF f used in Theorem 2.

To illustrate the important foregoing claim, we consider the following strategy profile σ where $\sigma_i(t_i) = (m_i^1, m_i^2, m_i^3, m_i^4)$ such that

item A $m_i^1[i] = t_i$ (i.e., each player announces her own type truthfully)

item B $m_1^1[2] = t_2$ and $m_2^1[1] = t_1$ (i.e., player 1 always announces t_2 as player 2's type and player 2 always announces t_1 as player 1's type in the first component of the message)

item C $m_1^2 = m_2^2 = 1$ (i.e., each player announces one in the second component of the message)

By Step 1 of the proof of Theorem 2, every rationalizable strategy profile induces Rule 1. By construction, the strategy profile σ induces Rule 1. In Step 3 of the proof of Theorem 2, each such $\sigma_i(t_i)$ is rationalizable. However, we argue that the strategy profile σ does not constitute a Bayesian equilibrium. If this were true, either player 1 of some type or player 2 of some type has a profitable deviation that triggers Rule 2-1. We indeed show that type t_1' of player 1 has a profitable deviation that triggers Rule 2-1.

Player 1 of type t_1 receives the following payoff under σ:

$$U_1(f\,|\,t_1) = 0.99 \times 4 + 0.01 \times 3 = 3.99.$$

Define $y : T_2 \longrightarrow \Delta(A)$ such that $y(t_2) = y(t_2') = 0.99 \times a + 0.01 \times b$. Then, we obtain

$$U_1(y\,|\,t_1) = 0.99U_1(a\,|\,t_1) + 0.01U_1(b\,|\,t_1) = 0.99 \times 4 + 0.01 \times 0.03 = 3.9603 < 3.99,$$

where $U_1(a\,|\,t_1) = 4$ and $U_1(b\,|\,t_1) = 0.03$. This implies that $y \in Y_1^*[t_1,f]$. Next, we compute

$$U_1(y\,|\,t_1') = 0.99U_1(a\,|\,t_1') + 0.01U_1(b\,|\,t_1') = 0.99 \times 4 + 0.01 \times 2 = 3.98 > 3 = U_1(f\,|\,t_1'),$$

Table 6. The payoff profiles in each type profile when alternative d is chosen in Example 6.

d	t_2	t_2'
t_1	3, 4	2, 0
t_1'	0, 3	3, 3
t_1''	0, 3	3, 3

Table 7. The payoff profiles in each type profile when alternative z is chosen in Example 6.

z	t_2	t_2'
t_1	4, 1	2, 0
t_1'	2, 2	5, 0
t_1''	2, 2	2, 0

where $U_1(a|t_1')=4$ and $U_1(b|t_1')=2$. Define $\hat{m}_1=(\hat{m}_1^1,\hat{m}_1^2,\hat{m}_1^3,\hat{m}_1^4)$ as being the same as $\sigma_1(t_1')$ except that we set $m_1^3[t_1]=y$ and $\hat{m}_1^2$ as an integer high enough. Then, $\hat{m}_1$ becomes type t_1''s profitable deviation that triggers Rule 2-1 where player 2 announces $m_2^1[1]=t_1$. This shows that σ is not an equilibrium.

8. Discussion of Other Issues

8.1. Well-Behaved Mechanisms

We say that a mechanism $\Gamma=((M)_{i\in I},g)$ is *well behaved* if every type t_i of every agent i has a best response to any conjecture $\lambda_i\in\Delta(T_{-i}\times M_{-i})$ such that $\text{marg}_{T_{-i}}\lambda_i=\pi_i(t_i)$.[21] Our canonical mechanism proposed in Theorem 2, as many other implementing mechanisms in the literature, fail to satisfy this definition. In contrast, in their analysis of robust implementation, both Bergemann and Morris [9] and Ollár and Penta [30] define each T_i as a compact set of payoff types on the real line and focus on the associated direct mechanisms, which are well-behaved according to our definition.[22]

In finite environments such as ours, finite mechanisms are to be anticipated. Finite mechanisms are clearly well behaved. However, Dutta and Sen [18] construct a finite environment in which the necessity of infinite mechanisms is established even for pure-strategy Bayesian implementation. This implies that the restriction to finite mechanisms is not as innocuous as one might think, and this is why we do not impose it in our search for general results in rationalizable implementation, even in finite-type settings.

Bergemann and Morris [7] show that, if an SCF f is implementable in interim rationalizable strategies by a finite mechanism, it satisfies IRM. Because IRM implies BM (Bergemann and Morris [7]), it follows that, if an SCF f is implementable in interim rationalizable strategies by a finite mechanism, it satisfies BM. Indeed, if the implementing mechanism is well behaved, then, because of the single-valuedness of the SCF, any selection of the rationalizable strategies correspondence forms a pure-strategy Bayesian equilibrium in the mechanism. Therefore, if an SCF is implementable in interim rationalizable strategies by a well-behaved mechanism, it is also implemented in Bayesian equilibrium by the same mechanism. Thus, within the class of well-behaved mechanisms, interim rationalizable implementation could be more restrictive than Bayesian implementation. This boils down to the difference between BM and IRM, because, as we argue in the next section, IRM is necessary for interim rationalizable implementation by well-behaved mechanisms. The difference between these two properties is well illustrated in Example 2, where the SCF f^ϵ violates IRM but satisfies BM in some circumstances.

For complete information environments, Chen et al. [14] characterizes rationalizable implementation by means of finite mechanisms when lotteries and transfers are allowed. The characterization is in terms of Maskin monotonicity*, a strengthening of Maskin monotonicity.[23] Chen et al. [15] also show, in the same environments, that Maskin monotonicity is necessary and sufficient for Nash implementation by finite mechanisms, thereby identifying a class of domains for which rationalizable implementation is more restrictive than Nash implementation. However, this result does not stand if one performs robust implementation: as shown in Kunimoto and Saran [23], using finite mechanisms, robust implementation in rationalizable strategies and in interim equilibria are equivalent.

8.2. Mechanisms with the Best Response Property

Bergemann and Morris [7] defines a class of (possibly countably infinite) mechanisms satisfying the best response property, which, assuming $S^{\Gamma(T)}$ is nonempty valued, includes well-behaved mechanisms as a special case. A mechanism $\Gamma=((M_i)_{i\in I},g)$ satisfies the *best response property* if, for each agent $i\in I$, there exists $\nu_i:T_{-i}\longrightarrow\Delta(M_{-i})$ such that

Table 8. The payoff profiles in each type profile when alternative z' is chosen in Example 6.

z'	t_2	t_2'
t_1	4, 0	4, 1
t_1'	2, 0	2, 2
t_1''	2, 0	5, 0

Table 9. The description of the SCF f in Example 6.

f	t_2	t_2'
t_1	a	b
t_1'	c	d
t_1''	c	d

$\nu_i(t_{-i})[m_{-i}] > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$ and for all $\tilde{t}_i \in T_i$,

$$\arg\max_{m_i \in M_i} \sum_{t_{-i}, m_{-i}} \pi_i(\tilde{t}_i)[t_{-i}]\nu_i(t_{-i})[m_{-i}]u_i(g(m_i, m_{-i}), (\tilde{t}_i, t_{-i})) \neq \emptyset.$$

Note that ν_i has to be defined independently of $\tilde{t}_i$.

Bergemann and Morris [7] shows that IRM is necessary for implementation in interim rationalizable strategies by a mechanism that satisfies the best response property. In its Proposition 5, Bergemann and Morris [7] claims that three conditions (BIC, IRM, and *no total indifference*) are sufficient for implementation in interim rationalizable strategies.[24] However, we are unable to replicate this proof. Specifically, we believe that the no total indifference condition is too weak to prove the claim.

If we strengthen no total indifference to the no-worst-rule (NWR) condition (Kunimoto [22]), then the following result is true.[25]

Theorem 3. *For any SCF f, if there exists an SCF $\hat{f} \approx f$ such that $\hat{f}$ satisfies IRM and the NWR condition, then the SCF f is implementable in interim rationalizable strategies by a mechanism that has a Bayesian equilibrium, and hence satisfies the best response property.*[26]

Thus, IRM is necessary and almost sufficient for implementation in interim rationalizable strategies by a mechanism satisfying the best response property. Because IRM implies BM and the best response property allows for countably infinite mechanisms with canonical type constructions, we might be tempted to conclude that implementation in interim rationalizable strategies is less permissive than Bayesian implementation. The relevant examples in Section 7 show that this conclusion is wrong.

The restriction to mechanisms satisfying the best response property is not innocuous because, for most SCFs, implementation in interim rationalizable strategies by a mechanism satisfying the best response property implies that the SCF is Bayesian implementable. To see this, pick any SCF f satisfying the NWR condition. If f is implementable in interim rationalizable strategies by a mechanism satisfying the best response property, then it must satisfy IRM. However, then Theorem 3 implies that f is implementable in interim rationalizable strategies by a mechanism that has a Bayesian equilibrium, and thus this mechanism must implement f in Bayesian equilibria. If there are only two agents, one can actually prove a stronger result: If an SCF is implementable in interim rationalizable strategies by a mechanism satisfying the best response property, then the same mechanism must implement the SCF in Bayesian equilibrium. In contrast, if an SCF satisfies weak IRM but not IRM (as in several examples in Section 7), the SCF could be implemented in rationalizable strategies, but the implementing mechanism cannot satisfy the best response property (thus, it cannot be finite or well behaved) because IRM is necessary for implementation by mechanisms satisfying that property.

8.3. Double Implementation

The foregoing discussion may lead to the question of double implementation in Bayesian equilibrium and rationalizable strategies. Let $B^{\Gamma(T)}$ be the set of (possibly mixed) Bayesian equilibria in the game $\Gamma(T)$. That is,

$$B^{\Gamma(T)} = \{\sigma \in \Sigma \mid \sigma \text{ constitutes a Bayesian equilibrium of the game } \Gamma(T)\},$$

where $\Sigma = \Sigma_1 \times \cdots \times \Sigma_n$ and $\Sigma_i = \{\sigma_i \mid \sigma_i : T_i \to \Delta(M_i)\}$. Recall that any message profile that is played by some types in a Bayesian equilibrium is rationalizable for those types. This leads to the following definition of double implementation.

Definition 12. A mechanism Γ *doubly implements* an SCF f in Bayesian equilibria and rationalizable strategies if there exists an SCF $\hat{f} \approx f$ such that the following two conditions hold:

1. Nonemptiness: $B^{\Gamma(T)} \neq \emptyset$.
2. Uniqueness: for any $t \in T$, $m \in S^{\Gamma(T)}(t)$ implies $g(m) = \hat{f}(t)$.

We now argue that IRM is necessary and almost sufficient for double implementation in Bayesian equilibria and rationalizable strategies. If a mechanism doubly implements an SCF in Bayesian equilibria and rationalizable strategies, then the mechanism must satisfy the best response property (due to the existence of Bayesian equilibrium).

Therefore, as per the necessity result of Bergemann and Morris [7], IRM must be necessary for double implementation. Moreover, Theorem 3 implies that IRM and the NWR condition are sufficient conditions for double implementation.

To get additional implications, we revisit Example 2. In that example, we show that there are some circumstances in which the SCF f^{ϵ} satisfies weak IRM and BM but violates IRM (Claim 3). This means that, in these circumstances, f^{ϵ} cannot be doubly implemented by a single mechanism in both Bayesian equilibrium and rationalizable strategies. There could, however, exist necessarily distinct mechanisms that achieve separate implementations of f^{ϵ} in interim rationalizable strategies and in Bayesian equilibrium.

8.4. Responsive SCFs

An SCF f is *responsive* if, for all $i \in I$ and $t_i, t_i' \in T_i$: $t_i \neq t_i' \Rightarrow t_i \nsim_i^f t_i'$. Otherwise, f is *nonresponsive*. Then, for a responsive SCF, we can establish that weak IRM and IRM are identical conditions.

Theorem 4. *If the SCF f is responsive, then f satisfies IRM if and only if it satisfies weak IRM.*

Proof. The key to the proof is based on the observation that SIRBIC and responsiveness imply strict BIC. Using the strict BIC of f, one can construct an SCF y^{ϵ} close to f with its desired property for f to satisfy IRM. The detailed proof is in Appendix A. □

Thus, within the class of responsive SCFs, there is no difference between IRM and weak IRM. This result helps us better delineate the boundary between Bayesian implementation and implementation in interim rationalizable strategies. Together with Theorem 3, the result implies that, within the class of responsive SCFs, there is essentially no gap between implementation in interim rationalizable strategies and Bayesian implementation.

Having said that, responsiveness could be a very restrictive requirement in many situations of interest to the designer. For example, the designer could be facing types who have the same ex post preferences but who differ in their beliefs. In such environments, any SCF that depends only on the agents' ex post preferences will naturally be nonresponsive. Even in environments where different types have different ex post preferences, there are important examples where the SCF is nonresponsive, for example, the second-best trading rule in bilateral trading. Hence, the class of nonresponsive SCFs is nontrivial and of interest in its own right.

8.5. Complete Information Environments

Multiple examples in Section 7 show that rationalizable implementation could be more permissive than equilibrium implementation. Interestingly, this relation reverses when one considers the restricted class of complete information environments and single-valued rules, that is, equilibrium implementation of SCFs is more permissive than rationalizable implementation. Bergemann et al. [11] shows that the necessary condition for rationalizable implementation is stronger than Maskin monotonicity, which is necessary for Nash implementation (Maskin [26]). The paper also gives an example of an SCF that is implementable in Nash equilibrium but not in rationalizable strategies. Xiong [45] provides a complete characterization of SCFs that are implementable in rationalizable strategies when there are at least three agents. For the sufficiency part of the argument, that paper constructs a mechanism in which the set of Nash equilibria is nonempty; therefore, the mechanism implements the SCF both in rationalizable strategies and Nash equilibrium. However, we emphasize that the restriction to SCFs is not innocuous in these results. Indeed, as shown in Kunimoto and Serrano [24] and Jain [21], when it comes to social choice correspondences, rationalizable implementation is more permissive than equilibrium implementation in complete information environments.

9. Concluding Remarks

We propose weak interim rationalizable monotonicity (weak IRM) as the key condition that almost characterizes, that is, it is necessary and almost sufficient, interim rationalizable implementation of social choice functions. We also show by means of examples that IRM and Bayesian monotonicity are *not* necessary for interim rationalizable implementation. This suggests that interim rationalizable implementation can be more permissive than Bayesian implementation. In obtaining this conclusion, it is important to consider mechanisms that are well behaved from the viewpoint of rationalizability (sets of rationalizable messages are nonempty) but lack Bayesian equilibria. Such mechanisms, as it turns out, are essential in uncovering the exact constraints that rationalizable play imposes on the decentralization of single-valued rules. One of the main contributions of our paper is to construct an implementing mechanism that works regardless of whether it has equilibria or not. Because the question we posed was the identification of conditions for rationalizable implementation with no restrictions on the mechanisms uses, we view this as an improvement over the canonical constructions in the existing literature. We also plan to explore generalizations of the current study's findings to multivalued social choice rules, that is, social choice sets, in a separate paper.

Acknowledgments

The authors thank the area editor, associate editor, and two anonymous referees for useful comments that lead to significant improvements in the paper and Georgy Artemov, Pierpaolo Battigalli, Antonio Penta, Arunava Sen, Yves Sprumont, Rajiv Vohra, and Siyang Xiong for helpful comments and encouragement. All remaining errors are our own.

Appendix A. Omitted Proofs

In this appendix, we provide the proofs we omitted in the main body of the paper.

Proof of Theorem 1. Suppose the mechanism $\Gamma = ((M_i)_{i\in I}, g)$ implements f in rationalizable strategies. Then, there exists an SCF $\hat{f} \approx f$ such that

1. Nonemptiness: $S_i^{\Gamma(T)}(t_i) \neq \emptyset$ for all $t_i \in T_i$ and $i \in I$.
2. Uniqueness: for any $t \in T$, $m \in S^{\Gamma(T)}(t)$ implies $g(m) = \hat{f}(t)$.

For any $i \in I$, $t_i \in T_i$, we set $m_i^{t_i} \in S_i^{\Gamma(T)}(t_i)$ (such a message $m_i^{t_i}$ exists by the nonemptyness requirement of implementability in interim rationalizable strategies). By the uniqueness requirement,

$$\hat{f}(t) = g(m_1^{t_1}, \dots, m_n^{t_n}),\ \forall t \in T.$$

We now argue that $\hat{f}$ satisfies weak IRM.

As $m_i^{t_i} \in S_i^{\Gamma(T)}(t_i)$, by the definition of rationalizable strategies, there exists a belief $\lambda_i^{t_i} \in \Delta(T_{-i} \times M_{-i})$ such that $\text{marg}_{T_{-i}} \lambda_i^{t_i} = \pi_i(t_i)$; $\lambda_i^{t_i}(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$; and

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

For each t_{-i} such that $\pi_i(t_i)[t_{-i}] > 0$, define the conditional distribution $\sigma_{-i}^{t_i}(t_{-i}) \in \Delta(M_{-i})$ as follows: for any $m_{-i} \in M_{-i}$,

$$\sigma_{-i}^{t_i}(t_{-i})[m_{-i}] = \frac{\lambda_i^{t_i}(t_{-i}, m_{-i})}{\pi_i(t_i)[t_{-i}]}.$$

For each t_{-i} such that $\pi_i(t_i)[t_{-i}] = 0$, let $\sigma_{-i}^{t_i}(t_{-i}) \in \Delta(M_{-i})$ denote the degenerate distribution that puts probability one on $m_{-i}^{t_{-i}}$, that is, $\sigma_{-i}^{t_i}(t_{-i})[m_{-i}^{t_{-i}}] = 1$. In either case, $\sigma_{-i}^{t_i}(t_{-i})[m_{-i}] > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$. This is true by construction if t_{-i} is such that $\pi_i(t_i)[t_{-i}] = 0$, whereas if t_{-i} is such that $\pi_i(t_i)[t_{-i}] > 0$, then $\sigma_{-i}^{t_i}(t_{-i})[m_{-i}] > 0 \Rightarrow \lambda_i^{t_i}(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$.

Now for each $m_i \in M_i$, define $y^{m_i, t_i} : T_{-i} \longrightarrow \Delta(A)$ as follows: for all $t_{-i} \in T_{-i}$,

$$y^{m_i, t_i}(t_{-i}) = \sum_{m_{-i} \in M_{-i}} \sigma_{-i}^{t_i}(t_{-i})[m_{-i}] g(m_i, m_{-i}).$$

Because $\text{marg}_{T_{-i}} \lambda_i^{t_i} = \pi_i(t_i)$, if $\pi_i(t_i)[t_{-i}] = 0$, then $\lambda_i^{t_i}(t_{-i}, m_{-i}) = 0$ for all $m_{-i} \in M_{-i}$. Hence,

$$\begin{aligned}
&\sum_{t_{-i}, m_{-i}} \lambda_i^{t_i}(t_{-i}, m_{-i}) u_i(g(m_i, m_{-i}), (t_i, t_{-i})) \\
&= \sum_{t_{-i}:\pi_i(t_i)[t_{-i}]>0} \sum_{m_{-i}} \lambda_i^{t_i}(t_{-i}, m_{-i}) u_i(g(m_i, m_{-i}), (t_i, t_{-i})) \\
&(\because \pi_i(t_i)[t_{-i}] = 0 \Rightarrow \lambda_i^{t_i}(t_{-i}, m_{-i}) = 0,\ \ \forall m_{-i}) \\
&= \sum_{t_{-i}:\pi_i(t_i)[t_{-i}]>0} \sum_{m_{-i}} \pi_i(t_i)[t_{-i}] \frac{\lambda_i^{t_i}(t_{-i}, m_{-i})}{\pi_i(t_i)[t_{-i}]} u_i(g(m_i, m_{-i}), (t_i, t_{-i})) \\
&= \sum_{t_{-i}:\pi_i(t_i)[t_{-i}]>0} \pi_i(t_i)[t_{-i}] \sum_{m_{-i}} \sigma_{-i}^{t_i}(t_{-i})[m_{-i}] u_i(g(m_i, m_{-i}), (t_i, t_{-i})) \\
&\left(\because \sigma_{-i}^{t_i}(t_{-i})[m_{-i}] = \frac{\lambda_i^{t_i}(t_{-i}, m_{-i})}{\pi_i(t_i)[t_{-i}]}\right) \\
&= \sum_{t_{-i}:\pi_i(t_i)[t_{-i}]>0} \pi_i(t_i)[t_{-i}] u_i(y^{m_i, t_i}(t_{-i}), (t_i, t_{-i})) \\
&(\because \text{by linearity of expected utility } u_i(\cdot, (t_i, t_{-i}))) \\
&= U_i(y^{m_i, t_i} | t_i).
\end{aligned} \tag{A.1}$$

Define the set

$$L_i(t_i) = \{y^{m_i, t_i} : m_i \in M_i\}.$$

Consider the message $m_i^{t_i}$ set forth in the beginning of the proof. Recall that $m_i^{t_i} \in S_i^{\Gamma(T)}(t_i)$. By the requirement of implementation and the fact that $\sigma_{-i}^{t_i}(t_{-i})[m_{-i}] > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$, we get

$$y^{m_i^{t_i}, t_i}(t_{-i}) = \hat{f}\,(t_i, t_{-i}),\ \forall t_{-i} \in T_{-i}.$$

Therefore, the following is true for all $m_i \in M_i$:

$$\begin{aligned} U_i(\hat{f}\,|t_i) = U_i(y^{m_i^{t_i}, t_i}|t_i) &= \sum_{t_{-i}, m_{-i}} \lambda_i^{t_i}(t_{-i}, m_{-i}) u_i(g(m_i^{t_i}, m_{-i}), (t_i, t_{-i})) \\ &\geq \sum_{t_{-i}, m_{-i}} \lambda_i^{t_i}(t_{-i}, m_{-i}) u_i(g(m_i, m_{-i}), (t_i, t_{-i})) \\ &= U_i(y^{m_i, t_i}|t_i), \end{aligned} \tag{A.2}$$

where the second and last equalities follow from (A.1), and the weak inequality follows because $m_i^{t_i}$ is a best response of type t_i against the belief $\lambda_i^{t_i}$.

We now claim that if m_i is such that $y^{m_i, t_i}(t_{-i}) \neq \hat{f}(t_i, t_{-i})$ for some $t_{-i} \in T_{-i}$, then it must be that

$$U_i(\hat{f}\ |t_i) > U_i(y^{m_i, t_i}|t_i).$$

If the foregoing strict inequality were not true, then it would follow from (A.2) that

$$\begin{aligned} & U_i(\hat{f}\,|t_i) = U_i(y^{m_i, t_i}|t_i) \\ \Rightarrow & \sum_{t_{-i}, m_{-i}} \lambda_i^{t_i}(t_{-i}, m_{-i}) u_i(g(m_i^{t_i}, m_{-i}), (t_i, t_{-i})) = \sum_{t_{-i}, m_{-i}} \lambda_i^{t_i}(t_{-i}, m_{-i}) u_i(g(m_i, m_{-i}), (t_i, t_{-i})). \end{aligned}$$

Thus, m_i would also be a best response of type t_i against the belief λ^{t_i} and hence $m_i \in S_i^{\Gamma(T)}(t_i)$. Then, by the requirement of implementation and the fact that $\sigma_{-i}^{t_i}(t_{-i})[m_{-i}] > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$, we get

$$y^{m_i, t_i}(t_{-i}) = \hat{f}\,(t_i, t_{-i}),\ \forall t_{-i} \in T_{-i},$$

which is a contradiction. This establishes that the previous strict inequality holds.

We are now ready to prove that $\hat{f}$ satisfies weak IRM. Consider any deception β. Define the message correspondence profile $S = (S_1, \ldots, S_n)$ such that

$$S_i(t_i) = \bigcup_{t_i' \in \beta_i(t_i)} S_i^{\Gamma(T)}(t_i').$$

Suppose β is unacceptable for $\hat{f}$ but not weakly refutable. Then, by definition of weak refutability, for all $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \nsim_i^{f} t_i$, there exists $\psi_i \in \Delta(T_{-i} \times T)$, which satisfies $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$, such that for all SCFs f' that satisfy $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$, we have

$$\sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(\hat{f}\,(t_i', \tilde{t}_{-i}), (t_i, t_{-i})) \geq \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})). \tag{A.3}$$

We first show that for any $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \sim_i^{\hat{f}} t_i$, there exists $\psi_i \in \Delta(T_{-i} \times T)$, which satisfies $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$, such that (A.3) holds for all SCFs f' that satisfy $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$.

Pick any $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \sim_i^{\hat{f}} t_i$. We set the belief $\psi_i \in \Delta(T_{-i} \times T)$ such that $\psi_i(t_{-i}, \tilde{t}) = 0$ whenever either $\tilde{t}_i \neq t_i$ or $\tilde{t}_{-i} \neq t_{-i}$ and $\psi_i(t_{-i}, \tilde{t}) = \pi_i(t_i)[t_{-i}]$ whenever $\tilde{t}_i = t_i$ and $\tilde{t}_{-i} = t_{-i}$. As $t_{-i} \in \beta_{-i}(t_{-i})$, the belief ψ_i satisfies $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$. Moreover, $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$.

Consider any SCF f' such that $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$. Then

$$\begin{aligned} \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(\hat{f}(t_i', \tilde{t}_{-i}), (t_i, t_{-i})) &= \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(\hat{f}(t_i, \tilde{t}_{-i}), (t_i, t_{-i})) \\ &= U_i(\hat{f}\,|t_i) \\ &\geq U_i(f'(t_i, \cdot)|t_i) \\ &= \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})), \end{aligned}$$

where the first equality follows from the fact that $t_i' \sim_i^{\hat{f}} t_i$, the second and last equalities follow from the construction of the belief ψ_i, and the inequality follows from the fact that $f'(t_i, \cdot) \in Y_i[t_i, \hat{f}]$.

Thus, if we combine the previous result with the hypothesis that β is not weakly refutable, then we can hypothesize that for all $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$, there exists $\psi_i \in \Delta(T_{-i} \times T)$, which satisfies $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$, such that (A.3) holds for all SCFs f' that satisfy $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$.[27]

We next show that $b(S) \geq S$. Pick any $i \in I$, $t_i \in T_i$, and $m_i' \in S_i(t_i)$. We now construct a belief $\lambda_i^\Gamma \in \Delta(T_{-i} \times M_{-i})$ satisfying $\lambda_i^\Gamma(t_{-i}, m_{-i}) > 0$ implies $m_{-i} \in S_{-i}(t_{-i})$ and $\text{marg}_{T_{-i}} \lambda_i^\Gamma = \pi_i(t_i)$ such that m_i' is a best response for agent i of type t_i against λ_i^Γ.

By the definition of S, we have $m_i' \in S_i^{\Gamma(T)}(t_i')$ for some $t_i' \in \beta_i(t_i)$. Then, by our hypothesis, there exists $\psi_i \in \Delta(T_{-i} \times T)$, which satisfies $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$, such that (A.3) holds for all SCFs f' that satisfy $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$.

Define the belief $\lambda_i^\Gamma \in \Delta(T_{-i} \times M_{-i})$ as follows: for any (t_{-i}, m_{-i}),

$$\lambda_i^\Gamma(t_{-i}, m_{-i}) = \sum_{\tilde{t}} \psi_i(t_{-i}, \tilde{t}) \times \sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}].$$

By construction, $\lambda_i^\Gamma(t_{-i}, m_{-i}) > 0$ implies that there exists $\tilde{t} \in T$ such that $\psi_i(t_{-i}, \tilde{t}) > 0$ and $\sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] > 0$. However, $\psi_i(t_{-i}, \tilde{t}) > 0$ implies $\tilde{t}_{-i} \in \beta_{-i}(t_{-i})$. Moreover, $\sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] > 0$ implies $m_{-i} \in S_{-i}^{\Gamma(T)}(\tilde{t}_{-i})$; recall the definition of $\sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}]$ from the beginning of this proof. Because $\tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $m_{-i} \in S_{-i}^{\Gamma(T)}(\tilde{t}_{-i})$, it follows from the definition of S that $m_{-i} \in S_{-i}(t_{-i})$.

Again, by construction, for all $t_{-i} \in T_{-i}$,

$$\text{marg}_{T_{-i}} \lambda_i^\Gamma(t_{-i}) = \sum_{m_{-i}} \lambda_i^\Gamma(t_{-i}, m_{-i}) = \sum_{\tilde{t}} \psi_i(t_{-i}, \tilde{t}) = \pi_i(t_i)[t_{-i}].$$

Thus, $\text{marg}_{T_{-i}} \lambda_i^\Gamma = \pi_i(t_i)$.

Pick any $\tilde{m}_i \in M_i$ and consider $y^{\tilde{m}_i, \tilde{t}_i}$ as defined earlier in the proof. Now define the SCF $f^{\tilde{m}_i}$ such that $f^{\tilde{m}_i}(\tilde{t}) = y^{\tilde{m}_i, \tilde{t}_i}(\tilde{t}_{-i})$ for all $\tilde{t} \in T$. Recall that if $\tilde{m}_i$ is such that $y^{\tilde{m}_i, \tilde{t}_i}(t_{-i}) \neq \hat{f}(\tilde{t}_i, t_{-i})$ for some $t_{-i} \in T_{-i}$, then it must be that $U_i(\hat{f} \mid \tilde{t}_i) > U_i(y^{\tilde{m}_i, \tilde{t}_i} \mid \tilde{t}_i)$. Therefore, $f^{\tilde{m}_i}(\tilde{t}_i, \cdot) = y^{\tilde{m}_i, \tilde{t}_i} \in Y_i[\tilde{t}_i, \hat{f}]$ for all $\tilde{t}_i \in T_i$. Therefore, Inequality (A.3) holds for $f^{\tilde{m}_i}$.

By the requirement of implementability, we have

$$\hat{f}(t_i', \tilde{t}_{-i}) = \sum_{m_{-i} \in M_{-i}} \sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] g(m_i', m_{-i}), \ \forall \tilde{t}_{-i} \in T_{-i}.$$

We are ready to show that m_i' is a best response for agent i of type t_i against λ_i^Γ. Consider any $\tilde{m}_i \in M_i$. Then

$$\begin{aligned}
&\sum_{t_{-i}, m_{-i}} \lambda_i^\Gamma(t_{-i}, m_{-i}) u_i(g(m_i', m_{-i}), (t_i, t_{-i})) \\
&= \sum_{t_{-i}, m_{-i}} \left(\sum_{\tilde{t}} \psi_i(t_{-i}, \tilde{t}) \times \sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] u_i(g(m_i', m_{-i}), (t_i, t_{-i})) \right) \\
&\quad \text{(by definition of } \lambda_i^\Gamma) \\
&= \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) \left(\sum_{m_{-i}} \sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] u_i(g(m_i', m_{-i}), (t_i, t_{-i})) \right) \\
&= \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i \left(\sum_{m_{-i}} \sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] g(m_i', m_{-i}), (t_i, t_{-i}) \right) \\
&\quad \text{(by linearity of expected utility } u_i(\cdot, (t_i, t_{-i}))) \\
&= \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(\hat{f}(t_i', \tilde{t}_{-i}), (t_i, t_{-i})) \\
&\quad \text{(by the requirement of implementability of } \hat{f}) \\
&\geq \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f^{\tilde{m}_i}(\tilde{t}), (t_i, t_{-i})) \\
&\quad (\because \text{ inequality (4) holds for } f^{\tilde{m}_i}) \\
&= \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(y^{\tilde{m}_i, \tilde{t}_i}(\tilde{t}_{-i}), (t_i, t_{-i})) \\
&\quad \text{(by definition of } f^{\tilde{m}_i}) \\
&= \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) \left(\sum_{m_{-i}} \sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] u_i(g(\tilde{m}_i, m_{-i}), (t_i, t_{-i})) \right) \\
&\quad \text{(by definition of } y^{\tilde{m}_i, \tilde{t}_i} \text{ and linearity of expected utility } u_i(\cdot, (t_i, t_{-i}))) \\
&= \sum_{t_{-i}, m_{-i}} \lambda_i^\Gamma(t_{-i}, m_{-i}) u_i(g(\tilde{m}_i, m_{-i}), (t_i, t_{-i}))
\end{aligned}$$

Because m_i' is a best response of player i of type t_i against λ_i^Γ satisfying $\lambda_i^\Gamma(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}(t_{-i})$ and $\text{marg}_{T_{-i}} \lambda_i^\Gamma = \pi_i(t_i)$, it follows by definition that $m_i' \in b_i(S)[t_i]$.

As $b(S) \geq S$, we have $S \leq S^{\Gamma(T)}$. Consider any $t \in T$ and $t' \in \beta(t)$. Pick a message profile $m^{t'} \in S^{\Gamma(T)}(t')$ as defined in the beginning of the proof. By definition, $g(m^{t'}) = \hat{f}(t')$. Now $S^{\Gamma(T)}(t') \subseteq S(t) \subseteq S^{\Gamma(T)}(t)$, where the first set inclusion follows from the definition of the message correspondence profile S and the second set inclusion follows from $S \leq S^{\Gamma(T)}$. Therefore, $m^{t'} \in S^{\Gamma(T)}(t)$. Hence, $g(m^{t'}) = \hat{f}(t)$ by the uniqueness requirement of implementation. Thus, $\hat{f}(t') = \hat{f}(t)$. Therefore, β is acceptable for $\hat{f}$, which is a contradiction. This completes the proof. □

Proof of Lemma 1. Suppose the SCF f satisfies weak IRM. Fix $i \in I$ and $t_i \in T_i$. Pick any $t_i' \in T_i$. If $t_i \sim_i^f t_i'$, then clearly $U_i(f|t_i) = U_i(f; t_i'|t_i)$.

Next, suppose $t_i \nsim_i^f t_i'$. Consider the deception β such that $\beta_j(t_j) = \{t_j\}$ for all $t_j \in T_j$ and $j \neq i$ but

$$\beta_i(\tilde{t}_i) = \begin{cases} \{t_i, t_i'\}, & \text{if } \tilde{t}_i = t_i \\ \{\tilde{t}_i\}, & \text{otherwise.} \end{cases}$$

Because $t_i \nsim_i^f t_i'$, the deception β is unacceptable for f. Hence, by weak IRM, it must be weakly refutable. That is, there exist $j \in I$, $\hat{t}_j \in T_j$, and $\hat{t}_j' \in \beta_j(\hat{t}_j)$ satisfying $\hat{t}_j' \nsim_j^f \hat{t}_j$ such that for any $\psi_j \in \Delta(T_{-j} \times T)$ satisfying $\psi_j(t_{-j}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-j} \in \beta_{-j}(t_{-j})$ and $\pi_j(\hat{t}_j)[t_{-j}] = \sum_{\tilde{t} \in T} \psi_j (t_{-j}, \tilde{t})$ for all $t_{-j} \in T_{-j}$, there exists an SCF f' such that $f'(\tilde{t}_j, \cdot) \in Y_j[\tilde{t}_j, f]$ for all $\tilde{t}_j \in T_j$ and

$$\sum_{t_{-j}, \tilde{t}} \psi_j(t_{-j}, \tilde{t}) u_j(f'(\tilde{t}), (\hat{t}_j, t_{-j})) > \sum_{t_{-j}, \tilde{t}} \psi_j(t_{-j}, \tilde{t}) u_j(f(\hat{t}_j', \tilde{t}_{-j}), (\hat{t}_j, t_{-j})).$$

Because $\hat{t}_j' \nsim_j^f \hat{t}_j$ and $\hat{t}_j' \in \beta_j(\hat{t}_j)$, it must be that $j = i$, $\hat{t}_j = t_i$, and $\hat{t}_j' = t_i'$.

Consider the belief ψ_i such that (i) $\psi_i(t_{-i}, \tilde{t}) = 0$ whenever either $\tilde{t}_i \neq t_i$ or $\tilde{t}_{-i} \neq t_{-i}$ and (ii) $\psi_i(t_{-i}, \tilde{t}) = \pi_i(t_i)[t_{-i}]$ whenever $\tilde{t}_i = t_i$ and $\tilde{t}_{-i} = t_{-i}$. As $t_{-i} \in \beta_{-i}(t_{-i})$, the belief ψ_i satisfies $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$. Moreover, $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$. Hence, we must have some SCF f' such that $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$ such that

$$\begin{aligned} U_i(f'|t_i) = \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})) &> \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f(t_i', t_{-i}), (t_i, t_{-i})) \\ &= U_i(f; t_i'|t_i). \end{aligned}$$

However, $f'(t_i, \cdot) \in Y_i[t_i, f]$ implies that $U_i(f|t_i) \geq U_i(f'|t_i)$. Therefore, $U_i(f|t_i) > U_i(f; t_i'|t_i)$, which completes the proof. □

Proof of Lemma 2. Pick any deception β that is unacceptable for an SCF f.

(Only-if part) Suppose f satisfies IRM. Then, there exist $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \nsim_i^f t_i$ such that for all $\phi_i \in \Delta(T_{-i} \times T_{-i})$ satisfying $\phi_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$, there exists $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such that

$$\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(y(\tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

We argue that the tuple (i, t_i, t_i') satisfies the requirement for strong refutability of β. Pick any belief $\psi_i \in \Delta(T_{-i} \times T)$ satisfying $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$.

Let $\phi_i' \in \Delta(T_{-i} \times T_{-i})$ be such that, for all $t_{-i}, \tilde{t}_{-i} \in T_{-i}$,

$$\phi_i'(t_{-i}, \tilde{t}_{-i}) = \sum_{\tilde{t}_i} \psi_i(t_{-i}, \tilde{t}_i, \tilde{t}_{-i}).$$

Then, by construction, $\phi_i'(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i'(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$. Therefore, it follows from IRM that there exists $y' \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such that

$$\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i'(t_{-i}, \tilde{t}_{-i}) u_i(y'(\tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i'(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})). \tag{A.4}$$

Define the SCF f' such that $f'(\tilde{t}) = y'(\tilde{t}_{-i})$ for all $\tilde{t} \in T$. Then $f'(\tilde{t}_i, \cdot) = y'$ for all $\tilde{t}_i$. Hence, f' is nonresponsive to agent i's type and $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i$. Moreover, it follows from (A.4) that

$$\begin{aligned} \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})) &= \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(y'(\tilde{t}_{-i}), (t_i, t_{-i})) \\ &= \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i'(t_{-i}, \tilde{t}_{-i}) u_i(y'(\tilde{t}_{-i}), (t_i, t_{-i})) \\ &> \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i'(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})) \\ &= \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})). \end{aligned}$$

Thus, β is strongly refutable.

(If-part) Suppose that every unacceptable deception for f is strongly refutable. Then, there exist $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \nsim_i^f t_i$ such that for all $\psi_i \in \Delta(T_{-i} \times T)$ satisfying $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$, there exists an SCF f' such that f' is nonresponsive to agent i's type, $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$, and

$$\sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

We argue that the tuple (i, t_i, t_i') satisfies the requirement in IRM for deception β. Pick any belief $\phi_i \in \Delta(T_{-i} \times T_{-i})$ satisfying $\phi_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$.

Let $\psi_i' \in \Delta(T_{-i} \times T)$ be such that $\psi_i'(t_{-i}, \tilde{t}) = 0$ whenever $\tilde{t}_i \neq t_i$ and $\psi_i'(t_{-i}, \tilde{t}) = \phi_i(t_{-i}, \tilde{t}_{-i})$ whenever $\tilde{t}_i = t_i$. Then, by construction, $\psi_i'(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i'(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$. Therefore, it follows from strong refutability of β that there exists an SCF f'' such that f'' is nonresponsive to agent i's type, $f''(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$, and

$$\sum_{t_{-i}, \tilde{t}} \psi_i'(t_{-i}, \tilde{t}) u_i(f''(\tilde{t}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}} \psi_i'(t_{-i}, \tilde{t}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})). \tag{A.5}$$

Define the mapping $y : T_{-i} \to \Delta(A)$ such that $y(\tilde{t}_{-i}) = f''(t_i, \tilde{t}_{-i})$ for all $\tilde{t}_{-i} \in T_{-i}$. Since f'' is nonresponsive to agent i's type, we have $y = f''(\tilde{t}_i, \cdot)$ for all $\tilde{t}_i$. Hence, $y = f''(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$. That is, $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$. Moreover, it follows from (A.5) that

$$\begin{aligned}
\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(y(\tilde{t}_{-i}), (t_i, t_{-i})) &= \sum_{t_{-i}, \tilde{t}} \psi_i'(t_{-i}, \tilde{t}) u_i(f''(\tilde{t}), (t_i, t_{-i})) \\
&> \sum_{t_{-i}, \tilde{t}} \psi_i'(t_{-i}, \tilde{t}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})) \\
&= \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).
\end{aligned}$$

Thus, f satisfies IRM. □

Proof of Lemma 3. We prove separate proofs of the two statements in the lemma.

We prove (a) first. Suppose an SCF f satisfies weak NWR. Pick any $i \in I$, $t_i \in T_i$, and $\phi_i \in \Delta(T_{-i} \times T_{-i})$.

First, it follows from the definition of weak NWR that there exists $\tilde{y} \in Y_i^w[t_i, f]$ such that $U_i(f | t_i) > U_i(\tilde{y} | t_i)$. To see this, consider the belief $\tilde{\phi}_i$ such that $\tilde{\phi}_i(t_{-i}, t_{-i}') = 0$ whenever $t_{-i}' \neq t_{-i}$ and $\tilde{\phi}_i(t_{-i}, t_{-i}') = \pi_i(t_i)[t_{-i}]$ whenever $t_{-i}' = t_{-i}$. Then, there must exist $\tilde{y}, \tilde{y}' \in Y_i^w[t_i, f]$ such that

$$\begin{aligned}
U_i(f | t_i) \geq U_i(\tilde{y}' | t_i) &= \sum_{t_{-i}, t_{-i}'} \tilde{\phi}_i(t_{-i}, t_{-i}') u_i(\tilde{y}'(t_{-i}'), (t_i, t_{-i})) \\
&> \sum_{t_{-i}, t_{-i}''} \tilde{\phi}_i(t_{-i}, t_{-i}') u_i(\tilde{y}(t_{-i}'), (t_i, t_{-i})) \\
&= U_i(\tilde{y} | t_i),
\end{aligned}$$

where the first weak inequality follows from the fact that $\tilde{y}' \in Y_i^w[t_i, f]$, and the strict inequality follows from weak NWR.

Second, because f satisfies weak NWR, there exist $y, y' \in Y_i^w[t_i, f]$ such that

$$\sum_{t_{-i}, t_{-i}'} \phi_i(t_{-i}, t_{-i}') u_i(y(t_{-i}'), (t_i, t_{-i})) > \sum_{t_{-i}, t_{-i}'} \phi_i(t_{-i}, t_{-i}') u_i(y'(t_{-i}'), (t_i, t_{-i})).$$

Pick any $\epsilon \in (0, 1)$ and define $y^\epsilon : T_{-i} \to \Delta(A)$ such that $y^\epsilon(t_{-i}) = (1 - \epsilon) y(t_{-i}) + \epsilon \tilde{y}(t_{-i})$ for all $t_{-i} \in T_{-i}$. We similarly define y'^ϵ. By construction, y^ϵ and y'^ϵ are such that

$$U_i(f | t_i) > U_i(y^\epsilon | t_i) \text{ and } U_i(f | t_i) > U_i(y'^\epsilon | t_i).$$

For ϵ sufficiently close to one, we have

$$\sum_{t_{-i}, t_{-i}'} \phi_i(t_{-i}, t_{-i}') u_i(y^\epsilon(t_{-i}'), (t_i, t_{-i})) > \sum_{t_{-i}, t_{-i}'} \phi_i(t_{-i}, t_{-i}') u_i(y'^\epsilon(t_{-i}'), (t_i, t_{-i})).$$

We fix any such sufficiently large ϵ.

Third, because $\Delta^*(A)$ is a dense subset of $\Delta(A)$, for each t_{-i}, there exists a sequence of lotteries $\{\ell^z(t_{-i})\}_{z=1}^\infty \in \Delta^*(A)$ converging to $y^\epsilon(t_{-i})$. For each $z \geq 1$, define $y^z : T_{-i} \to \Delta^*(A)$ such that $y^z(t_{-i}) = \ell^z(t_{-i})$ for all $t_{-i} \in T_{-i}$. Similarly, we can define $y'^z : T_{-i} \to \Delta^*(A)$ such that $y'^z(t_{-i})$ converges to $y'^\epsilon(t_{-i})$ for all $t_{-i} \in T_{-i}$. As T_{-i} is finite, there exists a sufficiently large integer z such that

$$U_i(f | t_i) > U_i(y^z | t_i) \text{ and } U_i(f | t_i) > U_i(y'^z | t_i)$$

and

$$\sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i})u_i(y^z(t'_{-i}),(t_i,t_{-i})) > \sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i})u_i(y'^z(t'_{-i}),(t_i,t_{-i})). \tag{A.6}$$

The first set of inequalities imply that $y^z, y'^z \in Y_i^*[t_i,f]$.

Last, because $y_i^{t_i,f}$, by construction, assigns a positive weight to all $y \in Y_i^*[t_i,f]$, if, contrary to what we want to establish, we had

$$\sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i})u_i(y_i^{t_i,f}(t'_{-i}),(t_i,t_{-i})) \geq \sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i})u_i(y(t'_{-i}),(t_i,t_{-i})),\ \forall y \in Y_i^*[t_i,f],$$

then it must be that

$$\sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i})u_i(y^z(t'_{-i}),(t_i,t_{-i})) = \sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i})u_i(y'^z(t'_{-i}),(t_i,t_{-i})),$$

which contradicts (A.6).

We prove (b) next. Suppose that an SCF f satisfies weak NWR. Pick any $i \in I$, $t_i \in T_i$, and $z_i^1 \in \Delta(T_{-i})$. As $\overline{\alpha}$ assigns a positive weight to all $a \in A$, if

$$\sum_{t_{-i}} z_i^1(t_{-i})u_i(\overline{\alpha},(t_i,t_{-i})) \geq \sum_{t_{-i}} z_i^1(t_{-i})u_i(a,(t_i,t_{-i})),\ \forall a \in A,$$

then it must be that

$$\sum_{t_{-i}} z_i^1(t_{-i})u_i(a,(t_i,t_{-i})) = \sum_{t_{-i}} z_i^1(t_{-i})u_i(a',(t_i,t_{-i})),$$

for all $a, a' \in A$. Now consider the belief $\tilde{\phi}_i \in \Delta(T_{-i} \times T_{-i})$ such that $\tilde{\phi}_i(t_{-i},t_{-i}) = z_i^1(t_{-i})$ for all $t_{-i} \in T_{-i}$. Then, by weak NWR, there must exist $\tilde{y}, \tilde{y}' \in Y_i^w[t_i,f]$ such that

$$\sum_{t_{-i},t'_{-i}} \tilde{\phi}_i(t_{-i},t'_{-i})u_i(\tilde{y}(t'_{-i}),(t_i,t_{-i})) > \sum_{t_{-i},t'_{-i}} \tilde{\phi}_i(t_{-i},t'_{-i})u_i(\tilde{y}'(t'_{-i}),(t_i,t_{-i})).$$

However, the left-hand side of the previous inequality equals $\sum_{t_{-i}} z_i^1(t_{-i})u_i(\tilde{y}(t_{-i}),(t_i,t_{-i}))$, whereas the right-hand side equals $\sum_{t_{-i}} z_i^1(t_{-i})u_i(\tilde{y}'(t_{-i}),(t_i,t_{-i}))$, which contradicts the fact that type t_i is indifferent over all alternatives when type t_i holds the belief z_i^1. □

Proof of Theorem 2. We propose the following mechanism $\Gamma = ((M_i)_{i\in I}, g)$ to prove the sufficiency result: For each individual i, pick any one type from T_i. We denote this type as t_i^*.

Each individual i sends a message $m_i = (m_i^1, m_i^2, m_i^3, m_i^4)$, where,

item A $m_i^1 = (m_i^1[j])_{j\in I}$ such that $m_i^1[j] \in T_j$ for all $j \in I$,

item B $m_i^2 \in \mathbb{N}$,

item C $m_i^3 = (m_i^3[t_i])_{t_i \in T_i}$ such that $m_i^3[t_i] \in Y_i^*[t_i,\hat{f}]$ for all $t_i \in T_i$,

item D and $m_i^4 \in A$.

Each M_i is countable.

The outcome function $g : M \to \Delta(A)$ is defined as follows: For each $m \in M$:

Rule 1: If $m_i^2 = 1$ for all $i \in I \Rightarrow g(m) = \hat{f}(m_1^1[1], m_2^1[2], \ldots, m_n^1[n])$.

Rule 2: If there exists $i \in I$ such that $m_i^2 > 1$ but $m_j^2 = 1$ for all $j \in I\backslash\{i\}$, then one of the following subrules apply.

Rule 2-1: If there exists $t_i \in T_i$ such that $m_j^1[i] = t_i$ for all $j \in I\backslash\{i\}$, then

$$g(m) = \begin{cases} m_i^3[t_i]((m_j^1[j])_{j\neq i}) & \text{with probability } m_i^2/(m_i^2+1), \\ y_i^{t_i,\hat{f}}((m_j^1[j])_{j\neq i}) & \text{with probability } 1/(m_i^2+1). \end{cases}$$

Rule 2-2: If $m_{j'}^1[i] \neq m_k^1[i]$ for some $j', k \in I\backslash\{i\}$, then

$$g(m) = \begin{cases} m_i^3[t_i^*]((m_j^1[j])_{j\neq i}) & \text{with probability } m_i^2/(m_i^2+1), \\ y_i^{t_i^*,\hat{f}}((m_j^1[j])_{j\neq i}) & \text{with probability } 1/(m_i^2+1). \end{cases}$$

Rule 3: In all other cases,

$$g(m) = \begin{cases} m_1^4 & \text{with probability } m_1^2/(1+m_1^2)n, \\ m_2^4 & \text{with probability } m_2^2/(1+m_2^2)n, \\ \vdots & \vdots \\ m_n^4 & \text{with probability } m_n^2/(1+m_n^2)n, \\ \overline{\alpha} & \text{with the remaining probability.} \end{cases}$$

We now prove that the mechanism Γ implements the SCF f in interim rationalizable strategies. The proof consists of Steps 1 through 3.

Step 1: $m_i \in S_i^{\Gamma(T)}(t_i) \Rightarrow m_i^2 = 1$.

Proof of Step 1. Suppose by way of contradiction that $m_i \in S_i^{\Gamma(T)}(t_i)$ but $m_i^2 > 1$. Then, m_i is a best response of individual i of type t_i against some conjecture $\lambda_i \in \Delta(T_{-i} \times M_{-i})$ satisfying $\text{marg}_{T_{-i}} \lambda_i = \pi_i(t_i)$.

For each $t_i' \neq t_i^*$ and $t'_{-i} \in T_{-i}$, we define

$$M^2_{-i}(t_i', t'_{-i}) = \{m_{-i} : m_j^2 = 1 \text{ and } m_j^1[i] = t_i', \ \forall j \neq i, \ \text{ and } \ (m_j^1[j])_{j \neq i} = t'_{-i}\}.$$

For t_i^* and each $t'_{-i} \in T_{-i}$, we define

$$M^2_{-i}(t_i^*, t'_{-i}) = \left\{ m_{-i} : \begin{array}{l} (m_j^1[j])_{j \neq i} = t'_{-i} \text{ and} \\ \text{either } m_j^2 = 1 \ \text{ and } \ m_j^1[i] = t_i^*, \ \forall j \neq i, \\ \text{or } m_j^2 = 1, \ \forall j \neq i, \text{ but } m_{j'}^1[i] \neq m_k^1[i] \ \text{ for some } \ j', k \neq i \end{array} \right\}.$$

Also, define

$$M^3_{-i} = \{m_{-i} : \text{there exist one or more } j \neq i \text{ such that } m_j^2 > 1\}.$$

Note that $((M^2_{-i}(\tilde{t}_i, t'_{-i}))_{\tilde{t}_i \in T_i, t'_{-i} \in T_{-i}}, M^3_{-i})$ defines a partition of M_{-i}. As $m_i^2 > 1$, if $m_{-i} \in M^2_{-i}(\tilde{t}_i, t'_{-i})$, then Rule 2 is used under the profile (m_i, m_{-i}), whereas if $m_{-i} \in M^3_{-i}$, then Rule 3 is used under the profile (m_i, m_{-i}).

For each $\tilde{t}_i \in T_i$, define

$$\Lambda_i^{2,\tilde{t}_i} = \sum_{t_{-i}, t''_{-i}} \ \sum_{m_{-i} \in M^2_{-i}(\tilde{t}_i, t''_{-i})} \lambda_i(t_{-i}, m_{-i}).$$

Thus, $\Lambda_i^{2,\tilde{t}_i}$ is the probability of the event that all other individuals report a message profile in $\cup_{t''_{-i}} M^2_{-i}(\tilde{t}_i, t''_{-i})$.

Also, define

$$\Lambda_i^3 = \sum_{t_{-i}} \ \sum_{m_{-i} \in M^3_{-i}} \lambda_i(t_{-i}, m_{-i}).$$

Thus, Λ_i^3 is the probability of the event that all other individuals report a message profile in M^3_{-i}.

If $\tilde{t}_i$ is such that $\Lambda_i^{2,\tilde{t}_i} > 0$, then define $\phi_i^{2,\tilde{t}_i} \in \Delta(T_{-i} \times T_{-i})$ such that for all $t_{-i}, t'_{-i} \in T_{-i}$,

$$\phi_i^{2,\tilde{t}_i}(t_{-i}, t'_{-i}) = \sum_{m_{-i} \in M^2_{-i}(\tilde{t}_i, t'_{-i})} \frac{\lambda_i(t_{-i}, m_{-i})}{\Lambda_i^{2,\tilde{t}_i}}.$$

Thus, $\phi_i^{2,\tilde{t}_i}(t_{-i}, t'_{-i})$ is the conditional probability of the event that the type profile of all other individuals is t_{-i}, and they report a message profile in $M^2_{-i}(\tilde{t}_i, t'_{-i})$ given the event that all other individuals report a message profile in $\cup_{t''_{-i}} M^2_{-i}(\tilde{t}_i, t''_{-i})$.

If the type profile of all other individuals is t_{-i} and those agents of types t_{-i} report a message profile in $M^2_{-i}(\tilde{t}_i, t'_{-i})$, then when individual i of type t_i plays m_i, she expects the outcome to be given by the lottery

$$\left(\frac{m_i^2}{1 + m_i^2}\right) m_i^3[\tilde{t}_i](t'_{-i}) + \left(1 - \frac{m_i^2}{1 + m_i^2}\right) y_i^{\tilde{t}_i, \hat{f}}(t'_{-i}).$$

As a result, conditional on the event that all other individuals report a message profile in $\cup_{t''_{-i}} M^2_{-i}(\tilde{t}_i, t''_{-i})$, the expected payoff of individual i of type t_i when playing m_i is

$$\begin{aligned} &\left(\frac{m_i^2}{1 + m_i^2}\right) \sum_{t_{-i}, t'_{-i}} \phi_i^{2,\tilde{t}_i}(t_{-i}, t'_{-i}) u_i(m_i^3[\tilde{t}_i](t'_{-i}), (t_i, t_{-i})) \\ &+ \left(1 - \frac{m_i^2}{1 + m_i^2}\right) \sum_{t_{-i}, t'_{-i}} \phi_i^{2,\tilde{t}_i}(t_{-i}, t'_{-i}) u_i(y_i^{\tilde{t}_i, \hat{f}}(t'_{-i}), (t_i, t_{-i})). \end{aligned} \tag{A.7}$$

If $\Lambda_i^3 > 0$, then define $\phi_i^3 \in \Delta(T_{-i})$ such that, for any $t_{-i} \in T_{-i}$,

$$\phi_i^3(t_{-i}) = \sum_{m_{-i} \in M^3_{-i}} \frac{\lambda_i(t_{-i}, m_{-i})}{\Lambda_i^3}.$$

Thus, $\phi_i^3(t_{-i})$ is the conditional probability of the event that the type profile of all other individuals is t_{-i}, and those agents of types t_{-i} report a message profile in M^3_{-i} given the event that all other individuals report a message profile in M^3_{-i}.

If the type profile of all other individuals is t_{-i} and those agents of types t_{-i} report a message profile $m_{-i} \in M^3_{-i}$, then when playing m_i, individual i of type t_i expects the outcome to be given by the lottery:

$$\frac{1}{n}\left(\frac{m_i^2}{1+m_i^2}\right)m_i^4 + \frac{1}{n}\left(1-\frac{m_i^2}{1+m_i^2}\right)\overline{\alpha} + \sum_{j\neq i}\left(\frac{1}{n}\left(\frac{m_j^2}{1+m_j^2}\right)m_j^4 + \frac{1}{n}\left(1-\frac{m_j^2}{1+m_j^2}\right)\overline{\alpha}\right).$$

As a result, conditional on the event that all other individuals report a message profile in M^3_{-i}, the expected payoff of individual i of type t_i when playing m_i is

$$\begin{aligned}&\frac{1}{n}\left(\frac{m_i^2}{1+m_i^2}\right)\sum_{t_{-i}}\phi_i^3(t_{-i})u_i(m_i^4,(t_i,t_{-i})) + \frac{1}{n}\left(1-\frac{m_i^2}{1+m_i^2}\right)\sum_{t_{-i}}\phi_i^3(t_{-i})u_i(\overline{\alpha},(t_i,t_{-i}))\\ &+\sum_{t_{-i}}\sum_{m_{-i}\in M^3_{-i}}\frac{\lambda_i(t_{-i},m_{-i})}{\Lambda_i^3}\sum_{j\neq i}\left(\frac{1}{n}\left(\frac{m_j^2}{1+m_j^2}\right)u_i(m_j^4,(t_i,t_{-i})) + \frac{1}{n}\left(1-\frac{m_j^2}{1+m_j^2}\right)u_i(\overline{\alpha},(t_i,t_{-i}))\right).\end{aligned} \tag{A.8}$$

Now let individual i of type t_i deviate to $\hat{m}_i = (m_i^1, \hat{m}_i^2, \hat{m}_i^3, \hat{m}_i^4)$ such that $\hat{m}_i^2 = m_i^2 + 1$. $\hat{m}_i^3$ is defined as follows: for each $\tilde{t}_i \in T_i$:

▷ If $\Lambda_i^{2,\tilde{t}_i} > 0$, then let $\hat{m}_i^3[\tilde{t}_i] \in Y_i^*[\tilde{t}_i, \hat{f}]$ be such that

$$\sum_{t_{-i},t'_{-i}}\phi_i^{2,\tilde{t}_i}(t_{-i},t'_{-i})u_i(\hat{m}_i^3[\tilde{t}_i](t'_{-i}),(t_i,t_{-i})) \geq \sum_{t_{-i},t'_{-i}}\phi_i^{2,\tilde{t}_i}(t_{-i},t'_{-i})u_i(m_i^3[\tilde{t}_i](t'_{-i}),(t_i,t_{-i}))$$

and

$$\sum_{t_{-i},t'_{-i}}\phi_i^{2,\tilde{t}_i}(t_{-i},t'_{-i})u_i(\hat{m}_i^3[\tilde{t}_i](t'_{-i}),(t_i,t_{-i})) > \sum_{t_{-i},t'_{-i}}\phi_i^{2,\tilde{t}_i}(t_{-i},t'_{-i})u_i(y_i^{\tilde{t}_i,\hat{f}}(t'_{-i}),(t_i,t_{-i})).$$

Such $\hat{m}_i^3[\tilde{t}_i]$ exists because of Lemma 3.

▷ If $\Lambda_i^{2,\tilde{t}_i} = 0$, then let $\hat{m}_i^3[\tilde{t}_i] = m_i^3[\tilde{t}_i]$.

- $\hat{m}_i^4$ is defined as follows:

▷ If $\Lambda_i^3 > 0$, then let $\hat{m}_i^4 \in A$ be such that

$$\sum_{t_{-i}}\phi_i^3(t_{-i})u_i(\hat{m}_i^4,(t_i,t_{-i})) \geq \sum_{t_{-i}}\phi_i^3(t_{-i})u_i(m_i^4,(t_i,t_{-i}))$$

and

$$\sum_{t_{-i}}\phi_i^3(t_{-i})u_i(\hat{m}_i^4,(t_i,t_{-i})) > \sum_{t_{-i}}\phi_i^3(t_{-i})u_i(\overline{\alpha},(t_i,t_{-i})).$$

Such $\hat{m}_i^4$ exists because of Lemma 3.

▷ If $\Lambda_i^3 = 0$, then let $\hat{m}_i^4 = m_i^4$.

If $\Lambda_i^{2,\tilde{t}_i} > 0$, then conditional on the event that all other individuals report a message profile in $\cup_{t''_{-i}} M^2_{-i}(\tilde{t}_i, t''_{-i})$, the expected payoff of individual i of type t_i when playing $\hat{m}_i$ is

$$\begin{aligned}&\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i},t'_{-i}}\phi_i^{2,\tilde{t}_i}(t_{-i},t'_{-i})u_i(\hat{m}_i^3[\tilde{t}_i](t'_{-i}),(t_i,t_{-i}))\\ &+\left(1-\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i},t'_{-i}}\phi_i^{2,\tilde{t}_i}(t_{-i},t'_{-i})u_i(y_i^{\tilde{t}_i,\hat{f}}(t'_{-i}),(t_i,t_{-i})),\end{aligned}$$

which is, by construction, greater than type t_i's expected payoff in (A.7) when she plays m_i.

If $\Lambda_i^3 > 0$, then conditional on the event that all other individuals report a message profile in M^3_{-i}, the expected payoff of individual i of type t_i when playing $\hat{m}_i$ is

$$\begin{aligned}&\frac{1}{n}\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i}}\phi_i^3(t_{-i})u_i(\hat{m}_i^4,(t_i,t_{-i})) + \frac{1}{n}\left(1-\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i}}\phi_i^3(t_{-i})u_i(\overline{\alpha},(t_i,t_{-i}))\\ &+\sum_{t_{-i}}\sum_{m_{-i}\in M^3_{-i}}\frac{\lambda_i(t_{-i},m_{-i})}{\Lambda_i^3}\sum_{j\neq i}\left(\frac{1}{n}\left(\frac{m_j^2}{1+m_j^2}\right)u_i(m_j^4,(t_i,t_{-i})) + \frac{1}{n}\left(1-\frac{m_j^2}{1+m_j^2}\right)u_i(\overline{\alpha},(t_i,t_{-i}))\right),\end{aligned}$$

which is, by construction, greater than type t_i's expected payoff in (A.8) when she plays m_i.

As $\sum_{\tilde{t}_i} \Lambda_i^{2,\tilde{t}_i} + \Lambda_i^3 = 1$ (because $m_i^2 > 1$), it follows that $\hat{m}_i$ is a better response for individual i of type t_i against λ_i, a contradiction. This completes the proof of Step 1. □

Step 2: For each $i \in I$ and $t_i \in T_i$, let

$$\beta_i(t_i) = \{t_i\} \cup \{t_i' \in T_i : \exists m_i \in S_i^{\Gamma(T)}(t_i) \text{ such that } m_i^1[i] = t_i'\}.$$

Then, the deception $\beta = (\beta_i)_{i \in I}$ is acceptable for $\hat{f}$.

Proof of Step 2: Suppose not, that is, β is unacceptable for $\hat{f}$. Then, by weak IRM, β must be weakly refutable. That is, there exist $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \overset{f}{\approx}_i t_i$ such that for all $\psi_i \in \Delta(T_{-i} \times T)$ satisfying $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$, there exists an SCF f' such that $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, \hat{f}]$ for all $\tilde{t}_i \in T_i$ and

$$\sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(\hat{f}(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

As $t_i' \overset{f}{\approx}_i t_i$ and $t_i' \in \beta_i(t_i)$, we can find a message $m_i \in S_i^{\Gamma(T)}(t_i)$ such that $m_i^1[i] = t_i'$. From Step 1, we know that $m_i^2 = 1$. Then, m_i is a best response to some belief $\lambda_i \in \Delta(T_{-i} \times M_{-i})$ such that $\lambda_i(t_{-i}, m_{-i}) > 0 \Rightarrow m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$ and $\text{marg}_{T_{-i}} \lambda_i = \pi_i(t_i)$. From Step 1, it follows that $\lambda_i(t_{-i}, m_{-i}) > 0$ implies $m_j^2 = 1$ for all $j \neq i$. We next define a partition of all those message profiles in M_{-i} such that $m_j^2 = 1$ for all $j \neq i$.

For each $\hat{t}_i \neq t_i^*$ and $\tilde{t}_{-i} \in T_{-i}$, we define

$$M_{-i}^1(\hat{t}_i, \tilde{t}_{-i}) = \{m_{-i} : m_j^2 = 1 \text{ and } m_j^1[i] = \hat{t}_i,\ \forall j \neq i, \text{and } (m_j^1[j])_{j \neq i} = \tilde{t}_{-i}\}.$$

For t_i^* and each $\tilde{t}_{-i} \in T_{-i}$, we define

$$M_{-i}^1(t_i^*, \tilde{t}_{-i}) = \left\{ m_{-i} : \begin{array}{l} (m_j^1[j])_{j \neq i} = \tilde{t}_{-i} \text{ and} \\ \text{either } m_j^2 = 1 \text{ and } m_j^1[i] = t_i^*,\ \forall j \neq i, \\ \text{or } m_j^2 = 1,\ \forall j \neq i, \text{ but } m_{j'}^1[i] \neq m_k^1[i] \text{ for some } j', k \neq i \end{array} \right\}.$$

Define the belief $\psi_i^1 \in \Delta(T_{-i} \times T)$ as follows: For each $t_{-i} \in T_{-i}$ and $\tilde{t} \in T$, let

$$\psi_i^1(t_{-i}, \tilde{t}) = \sum_{m_{-i} \in M_{-i}^1(\tilde{t}_i, \tilde{t}_{-i})} \lambda_i(t_{-i}, m_{-i}).$$

Thus, $\psi_i^1(t_{-i}, \tilde{t})$ is the probability of the event that the type profile of all other individuals is t_{-i} and those agents of types t_{-i} report a message profile in $M_{-i}^1(\tilde{t}_i, \tilde{t}_{-i})$. In this event, individual i of type t_i expects the outcome to equal $\hat{f}(t_i', \tilde{t}_{-i})$ when playing m_i. As a result, the expected payoff of individual i of type t_i when playing m_i is

$$\sum_{t_{-i}, \tilde{t}} \psi_i^1(t_{-i}, \tilde{t}) u_i(\hat{f}(t_i', \tilde{t}_{-i}), (t_i, t_{-i})). \tag{A.9}$$

Now, $\psi_i^1(t_{-i}, \tilde{t}) > 0$ implies that $\lambda_i(t_{-i}, m_{-i}) > 0$ for some $m_{-i} \in M_{-i}^1(\tilde{t}_i, \tilde{t}_{-i})$. However, $\lambda_i(t_{-i}, m_{-i}) > 0$ also implies that $m_{-i} \in S_{-i}^{\Gamma(T)}(t_{-i})$. Hence, because of the construction of β, we have $\tilde{t}_{-i} \in \beta_{-i}(t_{-i})$. Moreover, because $\lambda_i(t_{-i}, m_{-i}) > 0$ implies $m_j^2 = 1$ for all $j \neq i$, it follows that

$$\pi_i(t_i)[t_{-i}] = \sum_{m_{-i} \in M_{-i}} \lambda_i(t_{-i}, m_{-i}) = \sum_{m_{-i} \in \cup_{\tilde{t} \in T} M_{-i}^1(\tilde{t})} \lambda_i(t_{-i}, m_{-i}) = \sum_{\tilde{t} \in T} \psi_i^1(t_{-i}, \tilde{t}).$$

Therefore, it follows from weak refutability of β that there exists an SCF f' such that $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, \hat{f}]$ for all $\tilde{t}_i \in T_i$ and

$$\sum_{t_{-i}, \tilde{t}} \psi_i^1(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}} \psi_i^1(t_{-i}, \tilde{t}) u_i(\hat{f}(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

It is without loss of generality to assume that the SCF f' is such that $f'(\tilde{t}_i, \cdot) \in Y_i^*[\tilde{t}_i, \hat{f}]$ for all $\tilde{t}_i \in T_i$. To see this, pick any $\tilde{t}_i \in T_i$.

If $f'(\tilde{t}_i, \cdot) \in Y_i^*[\tilde{t}_i, \hat{f}]$, then for each integer $z \geq 1$ and $t_{-i} \in T_{-i}$, define $f^z(\tilde{t}_i, t_{-i}) = f'(\tilde{t}_i, t_{-i})$. Then $f^z(\tilde{t}_i, \cdot) \in Y_i^*[\tilde{t}_i, \hat{f}]$ for all z.

If $f'(\tilde{t}_i, \cdot) \notin Y_i^*[\tilde{t}_i, \hat{f}]$, then for each integer $z \geq 1$ and $t_{-i} \in T_{-i}$, define $f^z(\tilde{t}_i, t_{-i}) \in \Delta^*(A) \cup_{t_i' \in T_i} \{\hat{f}(t_i', t_{-i})\}$ such that (a) if $f'(\tilde{t}_i, t_{-i}) = \hat{f}(\tilde{t}_i, t_{-i})$, then $f^z(\tilde{t}_i, t_{-i}) = f'(\tilde{t}_i, t_{-i})$ for all z, whereas (b) if $f'(\tilde{t}_i, t_{-i}) \neq \hat{f}(\tilde{t}_i, t_{-i})$, then $f^z(\tilde{t}_i, t_{-i})$ converges to $f'(\tilde{t}_i, t_{-i})$ as $z \to \infty$. Because $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, \hat{f}]$ but $f'(\tilde{t}_i, \cdot) \notin Y_i^*[\tilde{t}_i, \hat{f}]$, it must be that $f'(\tilde{t}_i, t_{-i}) \neq \hat{f}(\tilde{t}_i, t_{-i})$ for some $t_{-i} \in T_{-i}$. This implies that $U_i(\hat{f} | \tilde{t}_i) > U_i(f'(\tilde{t}_i, \cdot) | \tilde{t}_i)$. As $f^z(\tilde{t}_i, \cdot)$ converges pointwise to $f'(\tilde{t}_i, \cdot)$, T_{-i} is finite, and $u_i(\cdot, t)$ is continuous over $\Delta(A)$, we can find a sufficiently large integer $\hat{z}[\tilde{t}_i]$ such that

$$U_i(\hat{f} | \tilde{t}_i) > U_i(f^{\hat{z}[\tilde{t}_i]}(\tilde{t}_i, \cdot) | \tilde{t}_i),\ \forall z > z[\tilde{t}_i].$$

Therefore, $f^{\hat{z}[\tilde{t}_i]}(\tilde{t}_i, \cdot) \in Y_i^*[\tilde{t}_i, \hat{f}]$ for all $z > z[\tilde{t}_i]$.

Consider the sequence of SCFs $\{f^z\}_{z\in\mathbb{N}}$ as defined previously. As f^z converges pointwise to f', T_i is finite, and $u_i(\cdot,t)$ is continuous over $\Delta(A)$, we can find a sufficiently large integer $\hat{z}$ such that $f^{\hat{z}}(\tilde{t}_i,\cdot)\in Y_i^*[\tilde{t}_i,\hat{f}]$ for all $\tilde{t}_i\in T_i$ and

$$\sum_{t_{-i},\tilde{t}}\psi_i^1(t_{-i},\tilde{t})u_i(f^{\hat{z}}\,(\tilde{t}),(t_i,t_{-i}))>\sum_{t_{-i},\tilde{t}}\psi_i^1(t_{-i},\tilde{t})u_i(\hat{f}(t_i',\tilde{t}_{-i}),(t_i,t_{-i})).$$

Therefore, $f'(\tilde{t}_i,\cdot)\in Y_i^*[\tilde{t}_i,\hat{f}]$ for all $\tilde{t}_i\in T_i$.

Now, let individual i of type t_i deviate to $\hat{m}_i=(m_i^1,\hat{m}_i^2,\hat{m}_i^3,m_i^4)$ such that

item A $\hat{m}_i^2>1$, where the specific value is chosen later.

item B $\hat{m}_i^3$ is defined as follows: $\hat{m}_i^3[\tilde{t}_i]=f'(\tilde{t}_i,\cdot)$ for all $\tilde{t}_i\in T_i$.

Consider the event that the type profile of all other individuals is t_{-i}, and those agents of types t_{-i} report a message profile in $M_{-i}^1(\tilde{t}_i,\tilde{t}_{-i})$. In this event, after the deviation to $\hat{m}_i$, type t_i of individual i expects the outcome to equal

$$\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)f'(\tilde{t}_i,\tilde{t}_{-i})+\left(1-\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)y_i^{\tilde{t}_i,\hat{f}}\,(\tilde{t}_{-i}).$$

As a result, the expected payoff of individual i of type t_i when deviating to $\hat{m}_i$ is

$$\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i},\tilde{t}}\psi_i^1(t_{-i},\tilde{t})u_i(f'(\tilde{t}),(t_i,t_{-i}))+\left(1-\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i},\tilde{t}}\psi_i^1(t_{-i},\tilde{t})u_i(y_i^{\tilde{t}_i,\hat{f}}\,(\tilde{t}_{-i}),(t_i,t_{-i})).$$

If $\hat{m}_i^2$ is large enough, then the previous expression is greater than type t_i's expected payoff in (A.9) when she plays m_i. It follows that $\hat{m}_i$ is a better response for individual i of type t_i against λ_i, a contradiction. Thus, β is acceptable. This completes the proof of Step 2. □

It follows from Steps 1 and 2 that $m\in S^{\Gamma(T)}(t)\Rightarrow g(m)=\hat{f}(t)$.

Step 3: Define the message correspondence profile $\mathcal{S}=(\mathcal{S}_1,\ldots,\mathcal{S}_n)$ where each $\mathcal{S}_i:T_i\longrightarrow 2^{M_i}$ such that for all $i\in I$ and $t_i\in T_i$:

$$\mathcal{S}_i(t_i)=\{(m_i^1,1,m_i^3,m_i^4):m_i^1[i]=t_i\}.$$

Then, we have $b(\mathcal{S})\geq\mathcal{S}$, which implies that $\mathcal{S}\leq S^{\Gamma(T)}$.

Proof of Step 3. Pick any $i\in I$, $t_i\in T_i$, and $m_i\in\mathcal{S}_i(t_i)$. Pick any $\tilde{\sigma}_{-i}:T_{-i}\longrightarrow M_{-i}$ such that, for all $j\neq i$ and $t_j\in T_j$, (i) $\tilde{\sigma}_j(t_j)\in\mathcal{S}_j(t_j)$ and (ii) $\tilde{\sigma}_j^1(t_j)[i]=t_i$. Let the belief $\lambda_i\in\Delta(T_{-i}\times M_{-i})$ be such that for all $t_{-i}\in T_{-i}$, $\lambda_i(t_{-i},m_{-i})=0$ whenever $m_{-i}\neq\tilde{\sigma}_{-i}(t_{-i})$. Then, by construction, $\lambda_i(t_{-i},m_{-i})>0$ implies that $m_{-i}\in\mathcal{S}_{-i}(t_{-i})$ and $\mathrm{marg}_{T_{-i}}\lambda_i=\pi_i(t_i)$. When holding the belief λ_i and playing m_i, individual i of type t_i expects the payoff of

$$\sum_{t_{-i}}\pi_i(t_i)[t_{-i}]u_i(\hat{f}\,(t_i,t_{-i}),(t_i,t_{-i})).$$

On the one hand, when deviating to $\hat{m}_i$ such that $\hat{m}_i^1[i]=t_i'$ and $\hat{m}_i^2=1$, then individual i of type t_i expects the payoff of

$$\sum_{t_{-i}}\pi_i(t_i)[t_{-i}]u_i(\hat{f}\,(t_i',t_{-i}),(t_i,t_{-i})),$$

which is not improving due to SIRBIC. Recall that weak IRM of $\hat{f}$ implies that $\hat{f}$ satisfies SIRBIC (Lemma 1). On the other hand, when deviating to $\hat{m}_i$ such that $\hat{m}_i^2>1$, then individual i of type t_i expects the payoff of

$$\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i}}\pi_i(t_i)[t_{-i}]u_i(\hat{m}_i^3[t_i](t_{-i}),(t_i,t_{-i}))+\left(1-\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i}}\pi_i(t_i)[t_{-i}]u_i(y_i^{t_i,\hat{f}}\,(t_{-i}),(t_i,t_{-i})).$$

As $\hat{m}_i^3[t_i]\in Y_i^*[t_i,\hat{f}]$, individual i of type t_i cannot improve the payoff by any such deviation. Hence, $m_i\in b_i(\mathcal{S})[t_i]$. This completes the proof of Step 3. □

Steps 1 through 3 together comprise the proof of the theorem. □

Proof of Proposition 1. Pick any unacceptable deception β. Then, there exist $i\in I$, $t_i\in T_i$, and $t_i'\in\beta_i(t_i)$ such that $t_i'\not\sim_i^f t_i$.

Fix any belief $\psi_i\in\Delta(T_{-i}\times T)$ such that $\psi_i(t_{-i},\tilde{t})>0\Rightarrow\tilde{t}_{-i}\in\beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}]=\sum_{\tilde{t}\in T}\psi_i(t_{-i},\tilde{t})$ for all $t_{-i}\in T_{-i}$. Pick any $\tilde{t}_i\in T_i$ such that $\sum_{t_{-i},\tilde{t}_{-i}}\psi_i(t_{-i},\tilde{t}_i,\tilde{t}_{-i})>0$.

First, suppose $\tilde{t}_i\not\sim_i^f t_i'$. As $t_i'\not\sim_i^f t_i$ and f satisfies COND-1, we can find $y\in Y_i^w[f,t_i']$ such that

$$u_i(y(\tilde{t}_{-i}),t_i)>u_i(f(t_i',\tilde{t}_{-i}),t_i),\ \forall\tilde{t}_{-i}\in T_{-i}.$$

For any $\epsilon\in[0,1]$, define $f^\epsilon(\tilde{t}_i,\cdot)=\epsilon y(\cdot)+(1-\epsilon)f(t_i',\cdot)$. Then the following holds for all $\epsilon>0$:

$$\sum_{t_{-i},\tilde{t}_{-i}}\psi_i(t_{-i},\tilde{t}_i,\tilde{t}_{-i})u_i(f^\epsilon(\tilde{t}_i,\tilde{t}_{-i}),t_i)>\sum_{t_{-i},\tilde{t}_{-i}}\psi_i(t_{-i},\tilde{t}_i,\tilde{t}_{-i})u_i(f(t_i',\tilde{t}_{-i}),t_i).$$

Because $\tilde{t}_i \nsim^f_i t'_i$, it follows from SIRBIC that

$$\sum_{\tilde{t}_{-i}} \pi_i[\tilde{t}_i](\tilde{t}_{-i}) u_i(f(\tilde{t}_i, \tilde{t}_{-i}), \tilde{t}_i) > \sum_{\tilde{t}_{-i}} \pi_i[\tilde{t}_i](\tilde{t}_{-i}) u_i(f(t'_i, \tilde{t}_{-i}), \tilde{t}_i).$$

Using the continuity of expected utility, there exists a sufficiently small but positive ϵ such that

$$\sum_{\tilde{t}_{-i}} \pi_i[\tilde{t}_i](\tilde{t}_{-i}) u_i(f(\tilde{t}_i, \tilde{t}_{-i}), \tilde{t}_i) > \sum_{\tilde{t}_{-i}} \pi_i[\tilde{t}_i](\tilde{t}_{-i}) u_i(f^{\epsilon}(\tilde{t}_i, \tilde{t}_{-i}), \tilde{t}_i).$$

Let $\epsilon(\tilde{t}_i)$ be any such sufficiently small but positive value of ϵ.

Second, suppose $\tilde{t}_i \nsim^f_i t'_i$. Then $t'_i \sim^f_i t_i$ implies that $\tilde{t}_i \nsim^f_i t_i$. Since f satisfies COND-1, we can find $y \in Y^w_i[\tilde{t}_i, f]$ such that

$$u_i(y(\tilde{t}_{-i}), t_i) > u_i(f(\tilde{t}_i, \tilde{t}_{-i}), t_i), \ \forall \tilde{t}_{-i} \in T_{-i}.$$

For any $\epsilon \in [0, 1]$, define $f^{\epsilon}(\tilde{t}_i, \cdot) = (1 - \epsilon) y(\cdot) + \epsilon f(t_i, \cdot)$. Because $\tilde{t}_i \nsim^f_i t_i$, it follows from SIRBIC that

$$\sum_{\tilde{t}_{-i}} \pi_i[\tilde{t}_i](\tilde{t}_{-i}) u_i(f(\tilde{t}_i, \tilde{t}_{-i}), \tilde{t}_i) > \sum_{\tilde{t}_{-i}} \pi_i[\tilde{t}_i](\tilde{t}_{-i}) u_i(f(t_i, \tilde{t}_{-i}), \tilde{t}_i).$$

Because $y \in Y^w_i[\tilde{t}_i, f]$, the following holds for all $\epsilon > 0$:

$$\sum_{\tilde{t}_{-i}} \pi_i[\tilde{t}_i](\tilde{t}_{-i}) u_i(f(\tilde{t}_i, \tilde{t}_{-i}), \tilde{t}_i) > \sum_{\tilde{t}_{-i}} \pi_i[\tilde{t}_i](\tilde{t}_{-i}) u_i(f^{\epsilon}(\tilde{t}_i, \tilde{t}_{-i}), \tilde{t}_i).$$

As $u_i(y(\tilde{t}_{-i}), t_i) > u_i(f(\tilde{t}_i, \tilde{t}_{-i}), t_i)$ for all $\tilde{t}_{-i} \in T_{-i}$, there exists sufficiently small but positive ϵ such that

$$\begin{aligned} \sum_{t_{-i}, \tilde{t}_{-i}} \psi_i(t_{-i}, \tilde{t}_i, \tilde{t}_{-i}) u_i(f^{\epsilon}(\tilde{t}_i, \tilde{t}_{-i}), t_i) &> \sum_{t_{-i}, \tilde{t}_{-i}} \psi_i(t_{-i}, \tilde{t}_i, \tilde{t}_{-i}) u_i(f(\tilde{t}_i, \tilde{t}_{-i}), t_i) \\ &= \sum_{t_{-i}, \tilde{t}_{-i}} \psi_i(t_{-i}, \tilde{t}_i, \tilde{t}_{-i}) u_i(f(t'_i, \tilde{t}_{-i}), t_i), \end{aligned}$$

where the equality follows from the fact that $\tilde{t}_i \sim^f_i t'_i$.

Let $\epsilon(\tilde{t}_i)$ be any such sufficiently small but positive value of ϵ. Now define the SCF f' as follows: for any $\tilde{t}_i \in T_i$ and $\tilde{t}_{-i} \in T_{-i}$,

$$f'(\tilde{t}_i, \tilde{t}_{-i}) = \begin{cases} f^{\epsilon(\tilde{t}_i)}(\tilde{t}_i, \tilde{t}_{-i}), & \text{if } \sum_{t_{-i}, \tilde{t}_{-i}} \psi_i(t_{-i}, \tilde{t}_i, \tilde{t}_{-i}) > 0, \\ f(\tilde{t}_i, \tilde{t}_{-i}), & \text{if } \sum_{t_{-i}, \tilde{t}_{-i}} \psi_i(t_{-i}, \tilde{t}_i, \tilde{t}_{-i}) = 0, \end{cases}$$

where $\epsilon(\tilde{t}_i)$ is as defined in the preceding arguments. By construction, $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$ and

$$\sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), t_i) > \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f(t'_i, \tilde{t}_{-i}), t_i).$$

It follows that f satisfies weak IRM. □

Proof of Proposition 2. We first show that f satisfies weak NWR. Pick any $i \in I$, $t_i \in T_i$, and $\phi_i \in \Delta(T_{-i} \times T_{-i})$. Fix $t'_{-i} \in T_{-i}$ arbitrarily. As $f(t_i, t'_{-i})$ is in the interior of $\Delta(A)$ and the environment satisfies NTI, we can find a lottery $y'(t'_{-i}) \in \Delta(A)$ such that $u_i(f(t_i, t'_{-i}), t_i) > u_i(y'(t'_{-i}), t_i)$. Then, $f(t_i, \cdot), y'(\cdot) \in Y^w_i[t_i, f]$ and

$$\sum_{t_{-i}, t'_{-i}} \phi_i(t_{-i}, t'_{-i}) u_i(f(t_i, t'_{-i}), t_i) \neq \sum_{t_{-i}, t'_{-i}} \phi_i(t_{-i}, t'_{-i}) u_i(y'(t'_{-i}), t_i).$$

Thus, f satisfies weak NWR.

Next, we show that f satisfies COND-1. Fix $i \in I$ and $t_i, t'_i \in T_i$ arbitrarily and suppose $t'_i \nsim^f_i t_i$. Pick any $t_{-i} \in T_{-i}$. To establish COND-1, we show the existence of a lottery $\ell(t_{-i}) \in \Delta(A)$ such that

$$u_i(f(t'_i, t_{-i}), t'_i) \geq u_i(\ell(t_{-i}), t'_i) \quad \text{and} \quad u_i(\ell(t_{-i}), t_i) > u_i(f(t'_i, t_{-i}), t_i).$$

This is so because if such $\ell(t_{-i})$ is constructed, we can set $y : T_{-i} \to \Delta(A)$ such that $y(t_{-i}) = \ell(t_{-i})$ for all $t_{-i} \in T_{-i}$. By construction of y, we confirm that $y \in Y^w_i[t'_i, f]$ and $u_i(y(t_{-i}), t_i) > u_i(f(t'_i, t_{-i}), t_i)$ for all $t_{-i} \in T_{-i}$. Thus, f satisfies COND-1.

Recall that, as A is countable, we denote it by $\{a_0, a_1, \ldots, a_k, \ldots\}$. For any $k \geq 1$, let

$$U_k = u_i(a_k, t_i) - u_i(a_0, t_i) \text{ and } U'_k = u_i(a_k, t'_i) - u_i(a_0, t'_i).$$

Because f is responsive only when preferences differ, $u_i(\cdot, t'_i)$ is not a positive affine transformation of $u_i(\cdot, t_i)$. Then, the vectors $(U'_k)_{k\geq 1}$ and $(U_k)_{k \geq 1}$ are not codirectional; that is, there does not exist an $\alpha > 0$ such that $U'_k = \alpha U_k$ for all $k \geq 1$. Using NTI, we can strengthen the previous statement to claim that there does not exist an $\alpha \geq 0$ such that $U'_k = \alpha U_k$ for all $k \geq 1$.

We next show that there exist lotteries $\ell', \ell'' \in \Delta(A)$ such that both ℓ' and ℓ'' have finite supports and

$$u_i(\ell'', t_i') > u_i(\ell', t_i') \quad \text{and} \quad u_i(\ell', t_i) > u_i(\ell'', t_i). \tag{A.10}$$

If A has only two elements, that is, $A = \{a_0, a_1\}$, then NTI implies that $U_1 \neq 0$ and $U_1' \neq 0$. Letting $\alpha = U_1'/U_1$, we have $U_1' = \alpha U_1$. Then it must be that $\alpha < 0$; otherwise, we will contradict the established claim that there does not exist an $\alpha \geq 0$ such that $U_1' = \alpha U_1$. Thus, if $U_1 > 0$, then (A.10) is true when $\ell' = a_1$ and $\ell'' = a_0$, whereas if $U_1 < 0$, then (A.10) is true when $\ell' = a_0$ and $\ell'' = a_1$.

Next, consider the case when A has three or more elements.

First, suppose there exists an $\alpha < 0$ such that $U_k' = \alpha U_k$ for all $k \geq 1$. Because of NTI, there exists $\hat{k} \geq 1$ such that $U_{\hat{k}} \neq 0$. Thus, if $U_{\hat{k}} > 0$, then (A.10) is true when $\ell' = a_{\hat{k}}$ and $\ell'' = a_0$, whereas if $U_{\hat{k}} < 0$, then (A.10) is true when $\ell' = a_0$ and $\ell'' = a_{\hat{k}}$.

Second, suppose there exists no α such that $U_k' = \alpha U_k$ for all $k \geq 1$. (This is the only option left because we have already established that there does not exist an $\alpha \geq 0$ such that $U_k' = \alpha U_k$ for all $k \geq 1$.) Because of NTI, there exists $\hat{k} \geq 1$ such that $U_{\hat{k}} \neq 0$. Let $\alpha = U_{\hat{k}}'/U_{\hat{k}}$. Because A has at least three elements, by our hypothesis that $U_k' \neq \alpha U_k$ for all $k \geq 1$ and α, there exists $k \geq 1$ with $k \neq \hat{k}$ such that

$$U_k' \neq \frac{U_{\hat{k}}'}{U_{\hat{k}}} U_k.$$

Hence, for all $\varepsilon > 0$, there exists $(r_1, r_2) \in \mathbb{R}^2_{++}$ such that $|0.25 - r_1| + |0.25 - r_2| < \varepsilon$ and

$$0.25U_k' + 0.25U_{\hat{k}}{}' > r_1 U_k' + r_2 U_{\hat{k}}' \text{ and } r_1 U_k + r_2 U_{\hat{k}} > 0.25U_k + 0.25U_{\hat{k}}. \tag{A.11}$$

If ε is sufficiently small, we can guarantee that $r_1 + r_2 \leq 1$. Take any such ε and define lottery ℓ'' as the one that assigns probability 0.25 each to a_k and $a_{\hat{k}}$ and probability 0.5 to a_1. Also, define lottery ℓ' as the one that assigns probability r_1 to a_k, r_2 to $a_{\hat{k}}$, and probability $1 - r_1 - r_2$ to a_1. Then, (A.10) follows from (A.11).

Define $\ell(t_{-i}) = f(t_i', t_{-i}) + \delta(\ell' - \ell'')$ where $\ell' - \ell'' \equiv (\ell'[a] - \ell''[a])_{a \in A}$. Because $f(t_i', t_{-i})$ is in the interior of $\Delta(A)$ and both ℓ' and ℓ'' have finite supports, we can find a sufficiently small but positive δ such that $\ell(t_{-i}) \in \Delta(A)$ and

$$u_i(f(t_i', t_{-i}), t_i') > u_i(\ell(t_{-i}), t_i') \quad \text{and} \quad u_i(\ell(t_{-i}), t_i) > u_i(f(t_i', t_{-i}), t_i).$$

This completes the proof of the proposition. □

Proof of Theorem 4. Pick a responsive SCF f. Clearly, if f satisfies IRM, then it satisfies weak IRM. We prove that if f satisfies weak IRM, then it must satisfy IRM.

Because weak IRM implies SIRBIC (Lemma 1) and f is responsive, it follows that f satisfies strict BIC, that is, $U_i(f \mid t_i) > U_i(f; t_i' \mid t_i)$ for all $i \in I$ and $t_i, t_i' \in T_i$ such that $t_i \neq t_i'$.

Pick any unacceptable deception β. Because of weak IRM, β must be weakly refutable. That is, there exist $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \not\sim_i^f t_i$ such that for all $\psi_i \in \Delta(T_{-i} \times T)$ satisfying $\psi_i(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$, there exists an SCF f' such that $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$ and

$$\sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f'(\tilde{t}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}} \psi_i(t_{-i}, \tilde{t}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

Consider any belief $\phi_i \in \Delta(T_{-i} \times T_{-i})$ such that $\phi_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$.

Let $\psi_i' \in \Delta(T_{-i} \times T)$ be such that $\psi_i'(t_{-i}, \tilde{t}) = 0$ whenever $\tilde{t}_i \neq t_i'$ and $\psi_i'(t_{-i}, \tilde{t}) = \phi_i(t_{-i}, \tilde{t}_{-i})$ whenever $\tilde{t}_i = t_i'$. Then, by construction, $\psi_i'(t_{-i}, \tilde{t}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t} \in T} \psi_i'(t_{-i}, \tilde{t})$ for all $t_{-i} \in T_{-i}$. Therefore, it follows from weak refutability of β that there exists an SCF f'' such that $f''(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$, and

$$\begin{aligned} &\sum_{t_{-i}, \tilde{t}} \psi_i'(t_{-i}, \tilde{t}) u_i(f''(\tilde{t}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}} \psi_i'(t_{-i}, \tilde{t}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})) \\ \Rightarrow &\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f''(t_i', \tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})). \end{aligned} \tag{A.12}$$

Pick any $\epsilon \in (0, 1)$ and define $y^\epsilon : T_{-i} \to \Delta(A)$ such that $y^\epsilon(\tilde{t}_{-i}) = (1 - \epsilon) f''(t_i', \tilde{t}_{-i}) + \epsilon f(t_i', \tilde{t}_{-i})$ for all $\tilde{t}_{-i} \in T_{-i}$. Because $f''(t_i', \cdot) \in Y_i[t_i', f]$, it follows that $y^\epsilon \in Y_i[t_i', f]$ for all ϵ. Because of strict BIC of f and finiteness of T_i, there exists a sufficiently large $\epsilon < 1$ such that $y^\epsilon \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \neq t_i'$. Thus, for a sufficiently large $\epsilon < 1$, we have $y^\epsilon \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$. Moreover, it follows from (A.12) that

$$\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(y^\epsilon(\tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

Thus, f satisfies IRM. □

Appendix B. Equivalence Between Our IRM and (Strict) Interim Rationalizable Monotonicity in Bergemann and Morris [7]

Bergemann and Morris [7] formulates interim rationalizable monotonicity (referred to as *BM's-IRM*. differently from us. It also defines a stricter condition, termed strict interim rationalizable monotonicity (referred to as *BM's-strict-IRM*). We now argue that both these conditions are equivalent to IRM as defined in this paper.

Recall that we define a deception as a profile of correspondences $\beta = (\beta_1, \ldots, \beta_n)$ such that $\beta_i : T_i \to 2^{T_i}$ and $t_i \in \beta_i(t_i)$ for all $t_i \in T_i$ and $i \in I$. That is, we restrict a deception to always include truthful reports. To present BM's-(strict)-IRM, we need to first remove that restriction.

Thus, we define a *weak deception* as a profile of correspondences $\beta^w = (\beta_1^w, \ldots, \beta_n^w)$ such that $\beta_i^w : T_i \to 2^{T_i} \backslash \{\emptyset\}$ for all $i \in I$. A weak deception β^w is *acceptable for an SCF f* if, for all $t, t' \in T$, $t' \in \beta^w(t) \Rightarrow f(t) = f(t')$; otherwise, β^w is *unacceptable for f*.

Definition B.1. An SCF *f* satisfies *BM's-IRM* if, for every weak deception β^w that is unacceptable for *f*, there exist $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i^w(t_i)$ such that for all $\phi_i \in \Delta(T_{-i} \times T_{-i})$ satisfying $\phi_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}^w(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$, there exists $y \in \cap_{\tilde{t}_i \in T_i} Y_i^w[\tilde{t}_i, f]$ such that

$$\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(y(\tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

The SCF satisfies *BM's-strict-IRM* if, in addition, $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$.

Clearly, BM's-strict-IRM implies BM's-IRM. We now show that both of these are equivalent to IRM. However, first, we point out that BM's-IRM implies SIRBIC.

Lemma B.1. *If the SCF f satisfies BM's-IRM, then f satisfies SIRBIC.*

Proof. We leave the the proof of this lemma to the reader as it is similar to the proof of Lemma 1. □

Lemma B.2. *IRM, BM's-IRM, and BM's-strict-IRM are all equivalent.*

Proof. We prove the lemma in two steps. First, we argue that IRM implies BM's-strict-IRM. Second, we argue that BM's-IRM implies IRM. The result follows because BM's-strict-IRM implies BM's-IRM.

IRM implies BM's-strict-IRM: Suppose *f* satisfies IRM. Pick any unacceptable weak deception β^w. Then define the deception β as follows: $\beta_i(t_i) = \{t_i\} \cup \beta_i^w(t_i)$, for all $t_i \in T_i$ and $i \in I$. Because β^w is unacceptable, it follows that β is unacceptable. Then, by IRM, there exist $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ satisfying $t_i' \not\approx_i^f t_i$ such that for all $\phi_i \in \Delta(T_{-i} \times T_{-i})$ satisfying $\phi_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$, there exists $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such that

$$\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(y(\tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

Because $t_i' \in \beta_i(t_i)$ and $t_i' \not\approx_i^f t_i$, it must be that $t_i' \in \beta_i^w(t_i)$. Pick any $\phi_i \in \Delta(T_{-i} \times T_{-i})$ satisfying $\phi_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}^w(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$. Because $\beta_{-i}^w(t_{-i}) \subseteq \beta_{-i}(t_{-i})$, for all $t_{-i} \in T_{-i}$, it follows that ϕ_i satisfies $\phi_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$. Hence, there must exist $y \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such that

$$\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(y(\tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

Thus, *f* satisfies BM's-strict-IRM.

BM's-IRM implies IRM: Suppose *f* satisfies BM's-IRM. Pick any unacceptable deception β. As any deception is a weak deception, it follows that β is also an unacceptable weak deception. Then by BM's-IRM, we can find an $i \in I$, $t_i \in T_i$, and $t_i' \in \beta_i(t_i)$ such that for all $\phi_i \in \Delta(T_{-i} \times T_{-i})$ satisfying $\phi_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$, there exists $y \in \cap_{\tilde{t}_i \in T_i} Y_i^w[\tilde{t}_i, f]$ such that

$$\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(y(\tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

For any $\epsilon \in (0, 1)$, define $y^\epsilon(t_{-i}) = (1 - \epsilon) y(t_{-i}) + \epsilon(\frac{1}{2} f(t_i, t_{-i}) + \frac{1}{2} f(t_i', t_{-i}))$, for all $t_{-i} \in T_{-i}$.

Now, if $t_i' \not\approx_i^f t_i$, then the following is true for all $\tilde{t}_i \in T_i$: either $\tilde{t}_i \not\approx_i^f t_i$ or $\tilde{t}_i \not\approx_i^f t_i'$. Then, from the facts that $\epsilon > 0$, $y \in Y_i^w[\tilde{t}_i, f]$, and *f* satisfies SIRBIC (Lemma B.1), it follows that $U_i(f | \tilde{t}_i) > U_i(y^\epsilon | \tilde{t}_i)$, for all $\tilde{t}_i \in T_i$. Thus, $y^\epsilon \in \cap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$. Furthermore, if ϵ is small enough, then y^ϵ will also satisfy the following inequality:

$$\sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(y^\epsilon(\tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \phi_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

Thus, if $t_i' \not\approx_i^f t_i$, then we can conclude that *f* satisfies IRM.

Therefore, what is left to argue is that $t_i' \not\approx_i^f t_i$. Suppose otherwise. Then $f(t_i', t_{-i}) = f(t_i, t_{-i})$, for all t_{-i}. Now consider the belief $\tilde{\phi}_i \in \Delta(T_{-i} \times T_{-i})$ such that $\tilde{\phi}_i(t_{-i}, t_{-i}') = \pi_i(t_i)[t_{-i}]$, if $t_{-i} = t_{-i}'$, and $\tilde{\phi}_i(t_{-i}, t_{-i}') = 0$, if $t_{-i} \neq t_{-i}'$. Recall that $t_{-i} \in \beta_{-i}(t_{-i})$ by the definition of β. Thus, $\tilde{\phi}_i(t_{-i}, \tilde{t}_{-i}) > 0 \Rightarrow \tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ and $\pi_i(t_i)[t_{-i}] = \sum_{\tilde{t}_{-i} \in T_{-i}} \tilde{\phi}_i(t_{-i}, \tilde{t}_{-i})$ for all $t_{-i} \in T_{-i}$. Hence, by BM's-IRM, there exists $\tilde{y} \in \cap_{\tilde{t}_i \in T_i} Y_i^w[\tilde{t}_i, f]$ such that

$$\sum_{t_{-i}, \tilde{t}_{-i}} \tilde{\phi}_i(t_{-i}, \tilde{t}_{-i}) u_i(\tilde{y}(\tilde{t}_{-i}), (t_i, t_{-i})) > \sum_{t_{-i}, \tilde{t}_{-i}} \tilde{\phi}_i(t_{-i}, \tilde{t}_{-i}) u_i(f(t_i', \tilde{t}_{-i}), (t_i, t_{-i})).$$

However, the previous inequality is equivalent to $U_i(\tilde{y} \mid t_i) > U_i(f \mid t_i)$, which contradicts $\tilde{y} \in Y_i^w[t_i, f]$. Hence, we must have $t_i' \nsim_i^f t_i$, which completes the proof. □

Endnotes

[1] Some authors refer to the former property as "common knowledge of rationality" and to the latter as the "rational-expectations assumption." We remain neutral about such issues of terminology.

[2] For belief-free games, Battigalli and Siniscalchi [5] defines Δ-rationalizability by imposing extra restrictions on the first-order beliefs, and Battigalli et al. [6] shows that (a suitably defined) Δ-rationalizability is equivalent to interim correlated rationalizability.

[3] Although the authors have recently taken it down from their webpages, this draft still deserves much credit for some of the concepts we use, as explained in the sequel. Important related treatments were previously given for the case of virtual or approximate implementation (Abreu and Matsushima [1]), with its robust counterparts (Bergemann and Morris [8], Artemov et al. [2]; the latter paper using Δ-rationalizability). The different conclusions reached in Bergemann and Morris [8] and Artemov et al. [2] can be traced back to the different results in the two papers by Serrano and Vohra [39, 40], explained by the issue of negligibility of types that cannot be distinguished by their interim preferences. A recent paper, Kunimoto and Saran [23], studies the robust version of the implementation notion we use here. Ollár and Penta [31, 32] provide results for robust implementation using direct mechanisms when the agents commonly believe that types are drawn from identical distributions.

[4] Our formulation of IRM differs from that in Bergemann and Morris [7]. However, the two conditions, as well as *strict* interim rationalizable monotonicity (strict IRM) also defined in Bergemann and Morris [7], are all equivalent (see Appendix B). Although Oury and Terceiux [33] formulate strict IRM differently from Bergemann and Morris [7], it can be shown that those are also equivalent (the argument is similar to the one used to prove Lemma B.2 in Appendix B). IRM can also be seen as an incomplete information analogue of robust monotonicity (Bergemann and Morris [10]).

[5] Oury and Tercieux [33] is mainly concerned with continuous partial Bayesian implementation. The paper shows that if an SCF is strictly continuously partially Bayesian implementable, then it must satisfy IRM. It follows from our results that strict continuous partial Bayesian implementation is even more difficult than interim rationalizable implementation. Di Tillio [17] shows that continuous implementation in interim rationalizable strategies is not more demanding than interim rationalizable implementation when the designer is restricted to use finite mechanisms. That is, if a finite mechanism implements an SCF in interim rationalizable strategies, then the same mechanism continuously implements the SCF in interim rationalizable strategies. It remains an open question whether Di Tillio's result extends to infinite mechanisms, such as the canonical mechanism that we construct to prove our sufficiency result.

[6] See also Jain [21], which follows the approach in Mezzetti and Renou [27] of implementation via supports.

[7] Kunimoto and Saran [23] come to a similar conclusion for robust implementation.

[8] The fact that almost all previous papers in this literature have dealt with SCFs, as well as the fact that the issues under consideration are likely to be quite different (as our preliminary study of set-valued rules suggests) justify the restriction to functions in the current paper.

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

[10] There are several reasons to consider stochastic SCFs. First, the stochastic SCFs provide a more general treatment because a deterministic SCF is a special case thereof, and the designer might be interested in implementing a stochastic SCF. Second, since randomness in players' beliefs is natural in the context of rationalizability, the implementing mechanism might as well be stochastic. To the extent that the resulting outcome function of the mechanism corresponds to the SCF, we find it natural to include stochastic SCFs in the analysis. Finally, we wish to follow the literature (Bergemann and Morris [7], Oury and Tercieux [33]) in this assumption.

[11] For instance, a complete information environment is given by a type space such that $T_i = T_j$ for all $i, j \in I$, and the beliefs of any type $t_i \in T_i$ of player $i \in I$ are such that $\pi_i(t_i)[t_{-i}] = 1$ if $t_j = t_i$ for all $j \neq i$. In this case, $T^* = \{t \in T : t_i = t_j, \ \forall i, j \in I\}$. Thus, at each state $t \in T^*$, every agent always believes with probability one that all other agents' types are also the ones corresponding to t. Note that T^* is not a singleton in a nontrivial complete information environment: the true state in T^* is common knowledge among the agents, but there are multiple states that the planner must consider.

[12] The notion of equivalent SCFs is discussed in Jackson [19].

[13] Unlike Dekel et al. [16], we do not have the payoff-relevant state space separately from the type space in our formulation of interim correlated rationalizability. We chose this specification to be consistent with most of the papers on implementation in incomplete information environments.

[14] For our necessity result, we require that $S_i^{\Gamma(\mathcal{T})}(t_i) \neq \emptyset$ for all t_i. For sufficiency, our implementing mechanism has the same property.

[15] We refer the reader to Section 7.1.1, where we provide an intuitive account for weak IRM in private values environments.

[16] As noted in Endnote 4, Bergemann and Morris [7] formulate IRM slightly differently from us but the two conditions are equivalent; see Appendix B.

[17] See Postlewaite and Schmeidler [36], Palfrey and Srivastava [34], and Jackson [19] for the necessity of BM for implementation in pure Bayesian equilibrium and Serrano and Vohra [41] and Kunimoto [22] for the necessity of mixed BM for implementation in mixed Bayesian equilibrium. Mixed BM is a strictly stronger condition than BM, as shown in Example 1 of Serrano and Vohra [41].

[18] Note that $T^* = T$ in this case. Hence, $f \approx \hat{f}$ if and only if $f(t) = \hat{f}(t)$ for all $t \in T$.

[19] Suppose $t_i' \sim_i^f t_i$. For any $\varepsilon \in [0, 1]$, define $y^\varepsilon : T_{-i} \to \Delta(A)$ as follows: $y^{\epsilon}(t_{-i}) = \varepsilon \bar{a}(i) + (1 - \varepsilon) f(t_i, t_{-i})$, for all t_{-i}. Because of strategy proofness, we have the following: for all $\varepsilon > 0$: $u_i(y^\varepsilon(t_{-i}), t_i) > u_i(f(t_i', t_{-i}), t_i)$, for all t_{-i}. Because $t_i' \sim_i^f t_i$, it follows from SIRBIC and finiteness of T_{-i} that there exists a sufficiently small but positive ε such that $U_i(f \mid t_i') > U_i(y^\varepsilon \mid t_i')$. Thus, we have found the required $y^\varepsilon \in Y_i^w[f, t_i']$, implying COND-1 holds.

[20] The full discussion of this example is available in our previous working paper version.

[21] Our definition of well-behaved mechanisms is a natural extension of "the best response property" proposed by Jackson et al. [20]. That definition says that every agent has a best response to every (pure) strategy profile of the other agents. The authors argue that in order for the "Nash" part of the solution to make sense, we should require the best response property (p. 482).

[22] Ollár and Penta [30] allow for general belief restrictions, subsuming the fixed type-space model as a special case. In the quasilinear setting, the authors provide a “moment condition” that is sufficient to generate transfers to implement a differentiable and responsive SCF in interim rationalizable strategies using the associated direct mechanism.

[23] This condition features in Bergemann et al. [11]—for the rationalizable implementation of SCFs, albeit allowing general mechanisms.

[24] The SCF f satisfies *no total indifference* if for all $i \in I$ and $t_i \in T_i$, there exist $y, y' \in \cap_{\tilde{t}_i \in T_i} Y_i^{w}[\tilde{t}_i, f]$ such that $U_i(y|t_i) \neq U_i(y'|t_i)$.

[25] Oury and Tercieux [33] proves a similar result for three or more players, while imposing an extra condition, stronger than NWR.

[26] The proof requires a slight modification of the mechanism constructed to prove Theorem 2; the mechanism is similar to the one constructed by Bergemann and Morris [7] in its proof of proposition 5, except that that mechanism is not countable because the players can report elements in the reward set and the set of lotteries, which are not necessarily countable. The detailed proof is available upon request.

[27] We are able to drop $t' \nsim_i^{\hat{f}} t_i$ as part of the qualification in the hypothesis.

References

[1] Abreu D, Matsushima H (1992) Virtual implementation in iteratively undominated strategies: Incomplete information. Working paper, Princeton University, Princeton, NJ.

[2] Artemov G, Kunimoto T, Serrano R (2013) Robust virtual implementation: Toward a reinterpretation of the Wilson doctrine. *J. Econom. Theory* 148(2):424–447.

[3] Barberà S (2011) Strategy-proof social choice. Arrow K, Sen A, Suzumura K, eds. *Handbook of Social Choice and Welfare*, vol. 2 (Elsevier, Amsterdam), 731–831.

[4] Barberà S, Jackson M (1995) Strategy-proof exchange. *Econometrica* 63(1):51–87.

[5] Battigalli P, Siniscalchi M (2003) Rationalization and incomplete information. *BE J. Theoretical Econom.* 3(1):1–46.

[6] Battigalli P, Di Tillio A, Grillo E, Penta A (2011) Interactive epistemology and solution concepts in games with asymmetric information. *BE J. Theoretical Econom.* 11(Advances).

[7] Bergemann D, Morris S (2008) Interim rationalizable implementation. Working paper, Yale University, New Haven, CT.

[8] Bergemann D, Morris S (2009) Robust virtual implementation. *Theoretical Econom.* 4(1):45–88.

[9] Bergemann D, Morris S (2009) Robust implementation in direct mechanisms. *Rev. Econom. Stud.* 76:1175–1204.

[10] Bergemann D, Morris S (2011) Robust implementation in general mechanisms. *Games Econom. Behav.* 71(2):261–281.

[11] Bergemann D, Morris S, Tercieux O (2011) Rationalizable implementation. *J. Econom. Theory* 146(3):1253–1274.

[12] Bernheim D (1984) Rationalizable strategic behavior. *Econometrica* 52(4):1007–1028.

[13] Brandenburger A, Dekel E (1987) Rationalizability and correlated equilibria. *Econometrica* 55(6):1391–1402.

[14] Chen Y-C, Kunimoto T, Sun Y, Xiong S (2021) Rationalizable implementation in finite mechanisms. *Games Econom. Behav.* 129:181–197.

[15] Chen Y-C, Kunimoto T, Sun Y, Xiong S (2022) Maskin meets Abreu and Matsushima. *Theoretical Econom.* 17(4):1683–1717.

[16] Dekel E, Fudenberg D, Morris S (2007) Interim correlated rationalizability. *Theoretical Econom.* 2(1):15–40.

[17] Di Tillio A (2011) A robustness result for rationalizable implementation. *Games Econom. Behav.* 72(1):301–305.

[18] Dutta B, Sen A (1994) Bayesian implementation: The necessity of infinite mechanisms. *J. Econom. Theory* 64(1):130–141.

[19] Jackson M (1991) Bayesian implementation. *Econometrica* 59:461–477.

[20] Jackson M, Palfrey T, Srivastava S (1994) Undominated Nash implementation in bounded mechanisms. *Games Econom. Behav.* 6(3):474–501.

[21] Jain R (2021) Rationalizable implementation of social choice correspondences. *Games Econom. Behav.* 127:47–66.

[22] Kunimoto T (2019) Mixed Bayesian implementation in general environments. *J. Math. Econom.* 82:247–263.

[23] Kunimoto T, Saran R (2020) Robust implementation in rationalizable strategies in general mechanisms. Working paper, Singapore Management University, Singapore.

[24] Kunimoto T, Serrano R (2019) Rationalizable implementation of correspondences. *Math. Oper. Res.* 44(4):1326–1344.

[25] Lipman B (1994) A note on the implications of common knowledge of rationality. *Games Econom. Behav.* 6(1):114–129.

[26] Maskin E (1999) Nash equilibrium and welfare optimality. *Rev. Econom. Stud.* 66:23–38.

[27] Mezzetti C, Renou L (2012) Implementation in mixed Nash equilibrium. *J. Econom. Theory* 147(6):2357–2375.

[28] Mizukami H, Wakayama T (2007) Dominant strategy implementation in economic environments. *Games Econom. Behav.* 60(2):307–325.

[29] Myerson R, Satterthwaite M (1983) Efficient mechanisms for bilateral trading. *J. Econom. Theory* 19(2):265–281.

[30] Ollár M, Penta A (2017) Full implementation and belief restrictions. *Amer. Econom. Rev.* 107(8):2243–2277.

[31] Ollár M, Penta A (2022) Efficient full implementation via transfers: Uniqueness and sensitivity in symmetric environments. *AEA Paper Proc.* 112:438–443.

[32] Ollár M, Penta A (2023) A network solution to robust implementation: The case of identical but unknown distributions. *Rev. Econom. Stud.* Forthcoming.

[33] Oury M, Tercieux O (2012) Continuous implementation. *Econometrica* 80(4):1605–1637.

[34] Palfrey T, Srivastava S (1989) Implementation with incomplete information in exchange economies. *Econometrica* 57(1):115–134.

[35] Pearce D (1984) Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52(4):1029–1050.

[36] Postlewaite A, Schmeidler D (1986) Implementation in differential information economies. *J. Econom. Theory* 39(1):14–33.

[37] Roth A, Postlewaite A (1977) Weak vs. strong dominance in a market with indivisible goods. *J. Math. Econom.* 4(2):131–137.

[38] Saijo T, Sjöström T, Yamato T (2007) Secure implementation. *Theoretical Econom.* 2(3):203–229.

[39] Serrano R, Vohra R (2001) Some limitations of virtual Bayesian implementation. *Econometrica* 69(3):785–792.

[40] Serrano R, Vohra R (2005) A characterization of virtual Bayesian implementation. *Games Econom. Behav.* 50(2):312–331.

[41] Serrano R, Vohra R (2010) Multiplicity of mixed equilibria in mechanisms: A unified approach to exact and approximate implementation. *J. Math. Econom.* 46(5):775–785.

[42] Shapley L, Scarf H (1974) On cores and indivisibility. *J. Math. Econom.* 1(1):23–37.

[43] Svensson L-G (1999) Strategy-proof allocation of indivisible goods. *Soc. Choice Welfare* 16(4):557–567.

[44] Thomson W (2016) Non-bossiness. *Soc. Choice Welfare* 46(3):665–696.

[45] Xiong S (2023) Rationalizable implementation of social choice functions: Complete characterization. *Theoretical Econom.* 18(1):197–230.