BROWN
Orlando Bravo Center
for Economic Research

Interim Rationalizable Implementation of Functions*

Bravo Working Paper # 2020-023

Takashi Kunimoto,† Rene Saran,‡ and Roberto Serrano§

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 that weak IRM is a necessary and almost sufficient condition for rationalizable implementation. We show by means of an example that interim rationalizable monotonicity (IRM), found in the literature, is strictly stronger than weak IRM as its name suggests, and that IRM is not necessary for rationalizable implementation, as had been previously claimed. The same example also demonstrates that Bayesian monotonicity, the key condition for full Bayesian implementation, is not necessary for rationalizable implementation. This implies that rationalizable implementation can be more permissive than Bayesian implementation: one can exploit the fact that there are no mixed Bayesian equilibria in the implementing mechanism.

JEL Classification: C72, D78, D82.

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 (1984), Pearce (1984), Brandenburger and Dekel (1987), Lipman (1994)). When players are rational and there is common belief among them that this is the case, they must find themselves

∗We thank Pierpaolo Battigalli for helpful comments and encouragement. All remaining errors are our own.
†School of Economics, Singapore Management University, Singapore; tkunimoto@smu.edu.sg
‡Department of Economics, University of Cincinnati, Cincinnati, OH, USA; rene_saran@uc.edu
§Department of Economics, Brown University, Providence, RI, USA; roberto_serrano@brown.edu.

Interim Rationalizable Implementation of Functions*

Takashi Kunimoto†, Rene Saran‡, and Roberto Serrano§

This Version: October 2020

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 that weak IRM is a necessary and almost sufficient condition for rationalizable implementation. We show by means of an example that interim rationalizable monotonicity (IRM), found in the literature, is strictly stronger than weak IRM as its name suggests, and that IRM is not necessary for rationalizable implementation, as had been previously claimed. The same example also demonstrates that Bayesian monotonicity, the key condition for full Bayesian implementation, is not necessary for rationalizable implementation. This implies that rationalizable implementation can be more permissive than Bayesian implementation: one can exploit the fact that there are no mixed Bayesian equilibria in the implementing mechanism.

JEL Classification: C72, D78, D82.

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 (1984), Pearce (1984), Brandenburger and Dekel (1987), Lipman (1994)). When players are rational and there is common belief among them that this is the case, they must find themselves

*We thank Pierpaolo Battigalli for helpful comments and encouragement. All remaining errors are our own.

†School of Economics, Singapore Management University, Singapore; tkunimoto@smu.edu.sg

‡Department of Economics, University of Cincinnati, Cincinnati, OH, USA; rene.saran@uc.edu

§Department of Economics, Brown University, Providence, RI, USA; roberto_serrano@brown.edu

playing rationalizable strategies, without necessarily imposing the additional assumption that their beliefs are correct, as is the case in an equilibrium.[1] Its extension to incomplete information, our concern in this paper, is the notion of interim correlated rationalizability, due to Dekel, Fudenberg, and Morris (2007), which will be defined in a later section.[2]

Despite the impressive effort made by implementation theorists in the 1980's and 1990's, 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 (2008)) 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 4.5 and 6.3. Weak IRM is a weakening of the interim rationalizable monotonicity (IRM) condition proposed in Bergemann and Morris (2008), which will be shown not to be necessary for rationalizable implementation (Example 7.1). We stress this point because Oury and Tercieux (2012) makes an incorrect claim that IRM is necessary for interim rationalizable implementation in their footnote 4. IRM – but not weak IRM – implies Bayesian monotonicity, a necessary condition for implementation in Bayesian equilibrium (Lemma 5.8).[4] Indeed, we show in Example 7.1 that weak IRM can be satisfied even when Bayesian monotonicity fails. Our results thus

[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]Battigalli and Siniscalchi (2003) defines Δ-rationalizability by imposing extra restrictions on the first-order beliefs, and Battigalli *et al.* (2011) shows that (a suitably defined) Δ-rationalizability is equivalent to interim correlated rationalizability.

[3]Important related answers were previously given for the case of virtual or approximate implementation (Abreu and Matsushima (1992)), with its robust counterparts (Bergemann and Morris (2009), Artemov, Kunimoto, and Serrano (2013) – the latter paper using Δ-rationalizability). The different conclusions reached in Bergemann and Morris (2009) and Artemov, Kunimoto, and Serrano (2013) can be traced back to the different results in the two papers by Serrano and Vohra (2001, 2005), explained by the issue of negligibility of types that cannot be distinguished by their interim preferences. A recent paper – Kunimoto and Saran (2020) – studies the robust version of the implementation notion we use here.

[4]Oury and Tercieux (2012) are mainly concerned with continuous partial Bayesian implementation. They show 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 (2011) shows that continuous interim implementation in 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.

demonstrate that rationalizable implementation may be more permissive than equilibrium implementation.

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 to results in complete information environments. In contrast to our finding, Bergemann, Morris, and Tercieux (2011) and Xiong (2018) show that rationalizable implementation of SCFs under complete information is more restrictive than equilibrium implementation. For set-valued rules, however, Kunimoto and Serrano (2019) come to the reverse conclusion that rationalizable implementation is generally more permissive than equilibrium implementation under complete information.[5] For general correspondences, Kunimoto and Serrano (2019) identifies uniform monotonicity, which is a weakening of the classic Maskin Monotonicity (Maskin (1999)) and which reduces to it in the case of SCFs, as a necessary and almost sufficient condition for rationalizable implementation. Since 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 (2019), will also make them rationalizably implementable. On the other hand, it is easy to find set-valued rules that are implementable in rationalizable strategies, but not in Nash equilibrium. Our results show that the permissiveness of rationalizable implementation, in comparison to equilibrium implementation, carries over to incomplete information environments but now even for SCFs.[6] 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 (Example 7.1 illustrates this point well). We plan to generalize the findings in Kunimoto and Serrano (2019) as well as those in the current study by a separate paper, posing the question of set-valued rules under incomplete information.

This paper is organized as follows. Section 2 presents preliminaries. Section 3 introduces our notion of implementation in interim rationalizable strategies. Weak IRM, as the necessary condition for rationalizable strategies, 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

[5]See also Jain (2020), which follows the approach in Mezzetti and Renou (2012) of implementation via supports.

[6]Kunimoto and Saran (2020) come to a similar conclusion for robust implementation.

are sufficient for interim rationalizable implementation. Section 7 features our important Example 7.1 to show that IRM and Bayesian monotonicity are not necessary for interim rationalizable implementation, and Section 8 discusses the issues of finite mechanisms and complete information environments. Section 9 concludes the paper with a few open questions. Some proofs are relegated to the Appendix.

2 Preliminaries

Let $I = \{1, \ldots, n\}$ denote the finite set of agents and T_i be a finite set of types of agent i. Let $T \equiv T_1 \times \cdots \times T_n$, and $T_{-i} \equiv T_1 \times \cdots \times T_{i-1} \times T_{i+1} \times \cdots \times T_n$.[7] 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 finite 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. We let $\Delta^*(A)$ be any countable dense subset of $\Delta(A)$. 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 $f : T \to \Delta(A)$. Let $T^* \subseteq T$ be such that

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

We interpret T^* as the set of states the designer cares about. 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 $t \in T^*$.

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* associated with an SCF f, i.e., a mechanism where $M_i = T_i$ for all i and $g = f$.

In the direct revelation 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', while all other agents truthfully announce their types, is defined as:

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

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

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, Fudenberg, and Morris (2007)) as a solution concept and investigate the implications of implementation in interim correlated "rationalizable" strategies.[8] We fix a mechanism $\Gamma = (M, g)$ 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 $\bar{S} = (\bar{S}_1, \ldots, \bar{S}_n)$, where $\bar{S}_i(t_i) = M_i$ 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 $t_i \in T_i$,

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

Observe that b is increasing by definition: i.e., $S \leq S' \Rightarrow b(S) \leq b(S')$. By Tarski's fixed-point theorem, there is a largest fixed point of b, which we label $S^{\Gamma(T)}$. 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 $\bar{S}$ – the largest element of the lattice – and iteratively applying the operator b. Let the message correspondence profile $S^{\Gamma(T),0} = \bar{S}$ and, for all $i \in I$, $t_i \in T_i$, $k \geq 1$, iteratively define,

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

[8] Unlike Dekel, Fudenberg, and Morris (2007), 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.

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, since 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.[9] 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 3.1. 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.

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

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

Remark: These set-valued deceptions have already been used in previous literature on interim rationalizable implementation (Bergemenn and Morris (2008), Oury and Tercieux (2012)). On the other hand, 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 4.2. A deception β is *unacceptable for an SCF* f if there exist $t \in T$ and $t' \in \beta(t)$ such that $f(t) \neq f(t')$; otherwise, β is *acceptable for* f.

Unacceptable deceptions are a concern for the designer since they undermine her 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

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

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, i.e.,

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

Otherwise, we say $t_i \not\sim_i^f t_i'$. 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 *unresponsive to agent i's type* if $t_i \sim_i^f t_i'$ for all $t_i, t_i' \in T_i$.

Definition 4.3. 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 $t_i' \not\sim_i^f t_i$ 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

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

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 (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 upon which type of player i's behavior they 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 $\bigcup_{\tilde{t}_i \in T_i}\{\sigma_{-i}^{\tilde{t}_i}(\beta_{-i}(\hat{t}_{-i}))\}$, 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. In equilibrium implementation in incomplete information environments, such as 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 $\sigma^*(\beta_{-i}(\hat{t}_{-i}))$, 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 unresponsive 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 $y \in \bigcap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$. This will be important to understand the difference with the previous condition proposed in the literature. We discuss this in the next section.

Definition 4.4. 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 her attempt of

implementing f. If she insisted on the deception being strongly refutable, then the SCF would satisfy interim rationalizable monotonicity (IRM), a stronger condition introduced in the literature (Bergemann and Morris (2008), Oury and Tercieux (2012)). In particular, it is claimed in Oury and Tercieux (2012, 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 4.5. *If an SCF f is implementable in interim rationalizable strategies, then there exists an SCF $\hat{f} \approx f$ that satisfies weak IRM.*

Proof. 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,

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

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

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

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 $m_{-i}^{t_{-i}}$, i.e., $\sigma_{-i}^{t_i}(t_{-i})[m_{-i}^{t_{-i}}] = 1$. 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 $y^{m_i,t_i} : T_{-i} \to \Delta(A)$ as follows: for all $t_{-i} \in T_{-i}$,

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

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

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

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 $m_i \in M_i$:

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

$$= U_i(y^{m_i,t_i}|t_i), \tag{2}$$

where the second and last equalities follow from (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 (2) that

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

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 strict inequality above 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 $t_i' \not\sim_i^f t_i$, 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

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

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 (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

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

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 above 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 (3) holds for all SCFs $f^{'}$ that satisfy $f^{'}(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, f]$ for all $\tilde{t}_i \in T_i$.[10]

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 (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 $\sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] > 0$. But $\psi_i(t_{-i}, \tilde{t}) > 0$ implies $\tilde{t}_{-i} \in \beta_{-i}(t_{-i})$. Moreover, $\sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] >$

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

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. Since $\tilde{t}_{-i} \in \beta_{-i}(t_{-i})$ 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 $U_i(\hat{f}|\tilde{t}_i) > U_i(y^{\tilde{m}_i, \tilde{t}_i}|\tilde{t}_i)$. So $f^{\tilde{m}_i}(\tilde{t}_i, \cdot) = y^{\tilde{m}_i, \tilde{t}_i} \in Y_i[\tilde{t}_i, \hat{f}]$ for all $\tilde{t}_i \in T_i$. So inequality (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

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

$$
\begin{aligned}
&\quad (\because \text{inequality (3) holds for } f^{\tilde{m}_i}) \\
&= \sum_{t_{-i},\tilde{t}} \psi_i(t_{-i},\tilde{t}) u_i\big(y^{\tilde{m}_i,\tilde{t}_i}(\tilde{t}_{-i}),(t_i,t_{-i})\big) \\
&\quad \text{(by definition of } f^{\tilde{m}_i}) \\
&= \sum_{t_{-i},\tilde{t}} \psi_i(t_{-i},\tilde{t}) \left(\sum_{m_{-i}} \sigma_{-i}^{\tilde{t}_i}(\tilde{t}_{-i})[m_{-i}] u_i\big(g(\tilde{m}_i,m_{-i}),(t_i,t_{-i})\big) \right) \\
&\quad \text{(by definition of } y^{\tilde{m}_i,\tilde{t}_i} \text{ and linearity of expected utility } u_i(\cdot,(t_i,t_{-i}))) \\
&= \sum_{t_{-i},m_{-i}} \lambda_i^{\Gamma}(t_{-i},m_{-i}) u_i\big(g(\tilde{m}_i,m_{-i}),(t_i,t_{-i})\big) \\
&\quad \text{(by definition of } \lambda_i^{\Gamma}).
\end{aligned}
$$

Since $m_i^{'}$ 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, $\hat{f}(t^{'}) = \hat{f}(t)$. So β is acceptable for $\hat{f}$, which is a contradiction. This completes the proof. □

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

$$U_i(f|t_i) \geq U_i(f;t_i^{'}|t_i), \forall t_i^{'} \in T_i$$

If these constraints are strict whenever $t_i \not\sim_i^f t_i^{'}$, then we say that f satisfies *strict-if-responsive Bayesian incentive compatibility (SIRBIC).*

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

Lemma 5.2. *If an SCF f satisfies weak IRM, then it satisfies SIRBIC.*

Proof. 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 $t_i \not\sim_i^f t_i^{'}$. Consider the deception β such that $\beta_j(t_j) = \{t_j\}$ for all $t_j \in T_j$ and $j \neq i$ but

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

Since $t_i \not\sim_i^f t_i^{'}$, 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 $\hat{t}_j^{'} \not\sim_j^f \hat{t}_j$ 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

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

Since $\hat{t}_j^{'} \not\sim_j^f \hat{t}_j$ 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

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

But $f^{'}(t_i, \cdot) \in Y_i[t_i, f]$ 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. □

As discussed in the previous section when we defined weak refutability, one can propose a stronger notion of refutability.

Definition 5.3. 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 unresponsive 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

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

Remark: Note how the SCF f' in the statement for strong refutability is required to be unresponsive 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 $y \in \bigcap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i,f]$ 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 (2008) requires the existence of such mappings $y \in \bigcap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i,f]$ in order to undermine unacceptable deceptions. Indeed, as we show next, interim rationalizable monotonicity is equivalent to strong refutability of every unacceptable deception.

Definition 5.4. An SCF f satisfies *interim rationalizable monotonicity (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 $y \in \bigcap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i,f]$ 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 5.5. *An SCF f satisfies IRM if and only if every deception β that is unacceptable for f is strongly refutable.*[11]

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

Corollary 5.6. *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 *unacceptable for an SCF f* if $f(\beta^s(t)) \neq f(t)$ for some $t \in T$; otherwise, β^s is *acceptable for f*.

Next, we introduce a necessary condition for full implementation in Bayesian equilibrium:

[11] Proof is relegated to the Appendix.

Definition 5.7. 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

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

while 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, since 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 significanly reduced. For IRM, but not for weak IRM, we can show the following implication:

Lemma 5.8. *If an SCF f satisfies IRM, it satisfies BM.*

Proof. Suppose that the SCF f satisfies IRM. Fix a single-valued deception β^s that is unacceptable for f. Define the "multi-valued" deception β such that $\beta_i(t_i) = \{t_i, \beta^s_i(t_i)\}$ for all $t_i \in T_i$ and $i \in I$. Since β^s is unacceptable, the deception β is also unacceptable. By IRM, 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 $y \in \bigcap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such that

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

For each $(t_{-i}, \tilde{t}_{-i}) \in T_{-i} \times T_{-i}$, we set

$$\phi_i(t_{-i}, \tilde{t}_{-i}) = \begin{cases} \pi_i(t_i)[t_{-i}], & \text{if } \tilde{t}_{-i} = \beta^s_{-i}(t_{-i}) \\ 0, & \text{if } \tilde{t}_{-i} \neq \beta^s_{-i}(t_{-i}). \end{cases}$$

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 $t_{-i} \in T_{-i}$. Moreover, since $t_i^{'} \in \beta_i(t_i)$ is such that $t_i^{'} \not\sim_i^f t_i$, it follows from construction of β that $t_i^{'} = \beta^s_i(t_i)$. Therefore, the above inequality becomes

$$\sum_{t_{-i}} \pi_i(t_i)[t_{-i}] u_i\big(y(\beta^s_{-i}(t_{-i})), (t_i, t_{-i})\big) > \sum_{t_{-i}} \pi_i(t_i)[t_{-i}] u_i\big(f(\beta^s_i(t_i), \beta^s_{-i}(t_{-i})), (t_i, t_{-i})\big),$$

which is equivalent to $U_i(y \circ \beta^s_{-i}|t_i) > U_i(f \circ \beta^s|t_i)$. In addition, $y \in \bigcap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ implies

that for any $\tilde{t}_i \in T_i$,

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

Hence, f satisfies BM. □

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 (NWR) (as discussed below, our definition is weaker than the one appearing in Kunimoto (2019)).

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

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

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 6.1. 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 (2019) also defines a "no worst rule" condition which is stronger than our definition. Kunimoto (2019) requires the existence of mappings y and y' in the set $\bigcap_{\tilde{t}_i \in T_i} Y_i^w[\tilde{t}_i, f]$ whereas we only require the existence of y and y' in the set $Y_i^w[t_i, f]$.

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

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

Note that $Y_i^*[t_i, f] \subseteq Y_i^w[t_i, f]$. Since T_{-i} is finite and $\Delta^*(A) \bigcup_{t'_i \in T_i} \{f(t'_i, t_{-i})\}$ 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, since A is countable, we denote it by $\{a_0, a_1, \ldots, a_k, \ldots\}$. Then, we define

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

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

The following lemma notes two important consequences of weak NWR (proof is relegated to the Appendix):

Lemma 6.2. *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

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

(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

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

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

Theorem 6.3. *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.*

Proof. 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

- $m_i^1 = (m_i^1[j])_{j \in I}$ such that $m_i^1[j] \in T_j$ for all $j \in I$,
- $m_i^2 \in \mathbb{N}$,

- $m_i^3 = (m_i^3[t_i])_{t_i \in T_i}$ such that $m_i^3[t_i] \in Y_i^*[t_i, \hat{f}]$ for all $t_i \in T_i$,
- and $m_i^4 \in A$.

Note that each M_i is countable.

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

Rule 1: $m_i^2 = 1$ 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 sub-rules 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

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

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

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

Rule 3: In all other cases:

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

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. 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

$$M_{-i}^2(t_i^{'}, t_{-i}^{'}) \;=\; \left\{ m_{-i} : m_j^2 = 1 \text{ and } m_j^1[i] = t_i^{'}, \forall j \neq i, \text{and } (m_j^1[j])_{j \neq i} = t_{-i}^{'} \right\}.$$

For t_i^* and each $t_{-i}^{'} \in T_{-i}$, we define

$$M_{-i}^2(t_i^*, t_{-i}^{'}) \;=\; \left\{ m_{-i} : \begin{array}{l} (m_j^1[j])_{j \neq i} = t_{-i}^{'} \text{ and} \\ \text{either } m_j^2 = 1 \text{ and } m_j^1[i] = t_i^*, \forall j \neq i, \\ \text{or } m_j^2 = 1, \forall j \neq i, \text{ but } m_{j^{'}}^1[i] \neq m_k^1[i] \text{ for some } j^{'}, k \neq i \end{array} \right\}.$$

Also, define

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

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

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

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

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

Also, define

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

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

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 $t_{-i}, t_{-i}^{'} \in T_{-i}$,

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

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 $M_{-i}^2(\tilde{t}_i, t_{-i}^{'})$ given the event that all other individuals report a message profile in $\bigcup_{t_{-i}^{''}} M_{-i}^2(\tilde{t}_i, t_{-i}^{''})$.

If the type profile of all other individuals is t_{-i} and they report a message profile in

$M^2_{-i}(\tilde{t}_i, t'_{-i})$, then when individual i of type t_i plays m_i, she expects the outcome to be given by the lottery

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

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

$$\begin{aligned}&\left(\frac{m_i^2}{1+m_i^2}\right) \sum_{t_{-i}, t'_{-i}} \phi_i^{2,\tilde{t}_i}(t_{-i}, t'_{-i}) u_i\big(m_i^3[\tilde{t}_i](t'_{-i}), (t_i, t_{-i})\big) \\ &+ \left(1 - \frac{m_i^2}{1+m_i^2}\right) \sum_{t_{-i}, t'_{-i}} \phi_i^{2,\tilde{t}_i}(t_{-i}, t'_{-i}) u_i\big(y_i^{\tilde{t}_i,\hat{f}}(t'_{-i}), (t_i, t_{-i})\big). \qquad (4)\end{aligned}$$

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 they 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 they report a message profile $m_{-i} \in M^3_{-i}$, then when individual i of type t_i plays m_i, she expects the outcome to be given by the lottery

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

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 she plays m_i is

$$\begin{aligned}&\frac{1}{n}\left(\frac{m_i^2}{1+m_i^2}\right) \sum_{t_{-i}} \phi_i^3(t_{-i}) u_i\big(m_i^4, (t_i, t_{-i})\big) + \frac{1}{n}\left(1 - \frac{m_i^2}{1+m_i^2}\right) \sum_{t_{-i}} \phi_i^3(t_{-i}) u_i\big(\bar{\alpha}, (t_i, t_{-i})\big) \\ &+ \sum_{t_{-i}} \sum_{m_{-i} \in M^3_{-i}} \frac{\lambda_i(t_{-i}, m_{-i})}{\Lambda_i^3} \sum_{j \neq i}\left(\frac{1}{n}\left(\frac{m_j^2}{1+m_j^2}\right) u_i\big(m_j^4, (t_i, t_{-i})\big) + \frac{1}{n}\left(1 - \frac{m_j^2}{1+m_j^2}\right) u_i\big(\bar{\alpha}, (t_i, t_{-i})\big)\right). \qquad (5)\end{aligned}$$

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

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

  and

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

  Note that such $\hat{m}_i^3[\tilde{t}_i]$ exists because of Lemma 6.2.

  ▷ 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

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

  and

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

  Note that such $\hat{m}_i^4$ exists because of Lemma 6.2.

  ▷ 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 $\bigcup_{t''_{-i}} M_{-i}^2(\tilde{t}_i, t''_{-i})$, the expected payoff of individual i of type t_i when she plays $\hat{m}_i$ is

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

which is, by construction, greater than her expected payoff in (4) 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_{-i}^3, the expected payoff of individual i of type t_i when she plays $\hat{m}_i$ is

$$\frac{1}{n}\left(\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i}}\phi_i^3(t_{-i})u_i\big(\hat{m}_i^4,(t_i,t_{-i})\big)+\frac{1}{n}\left(1-\frac{\hat{m}_i^2}{1+\hat{m}_i^2}\right)\sum_{t_{-i}}\phi_i^3(t_{-i})u_i\big(\bar{\alpha},(t_i,t_{-i})\big)$$
$$+\sum_{t_{-i}}\sum_{m_{-i}\in M_{-i}^3}\frac{\lambda_i(t_{-i},m_{-i})}{\Lambda_i^3}\sum_{j\neq i}\left(\frac{1}{n}\left(\frac{m_j^2}{1+m_j^2}\right)u_i\big(m_j^4,(t_i,t_{-i})\big)+\frac{1}{n}\left(1-\frac{m_j^2}{1+m_j^2}\right)u_i\big(\bar{\alpha},(t_i,t_{-i})\big)\right),$$

which is, by construction, greater than her expected payoff in (5) 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. 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 $t_i^{'}\not\sim_i^f t_i$ 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 $f^{'}(\tilde{t}_i,\cdot)\in Y_i[\tilde{t}_i,\hat{f}]$ for all $\tilde{t}_i\in T_i$ and

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

As $t_i^{'}\not\sim_i^f t_i$ 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

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

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 they 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 she plays m_i. As a result, the expected payoff of individual i of type t_i when she plays m_i is

$$\sum_{t_{-i}, \tilde{t}} \psi_i^1(t_{-i}, \tilde{t}) u_i\big(\hat{f}(t_i', \tilde{t}_{-i}), (t_i, t_{-i})\big). \tag{6}$$

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

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

So, it follows from weak refutability of β that there exists and 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

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

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 $f^z(\tilde{t}_i, t_{-i}) \in \Delta^*(A) \bigcup_{t_i' \in T_i} \{\hat{f}(t_i', t_{-i})\}$ 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 $z \to \infty$. Since $f'(\tilde{t}_i, \cdot) \in Y_i[\tilde{t}_i, \hat{f}]$ 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 above. 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

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

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

- $\hat{m}_i^2 > 1$, where the specific value is chosen later.
- $\hat{m}_i^3$ 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 they 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 she deviates to $\hat{m}_i$ is

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

If $\hat{m}_i^2$ is large enough, then the above expression is greater than her expected payoff in (6) 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 $\mathcal{S}_i : T_i \to 2^{M_i}$ 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. Pick any $i \in I$, $t_i \in T_i$, and $m_i \in \mathcal{S}_i(t_i)$. Pick any $\tilde{\sigma}_{-i} : T_{-i} \to M_{-i}$ 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 $\text{marg}_{T_{-i}} \lambda_i = \pi_i(t_i)$. When individual i of type t_i holds the belief λ_i and plays m_i, then she expects the payoff of

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

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

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

which is not improving due to SIRBIC. Recall that weak IRM of $\hat{f}$ implies that $\hat{f}$ satisfies SIRBIC (Lemma 5.2). On the other hand, if she deviates to $\hat{m}_i$ such that $\hat{m}_i^2 > 1$, then she expects the payoff of

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

As $\hat{m}_i^3[t_i] \in Y_i^*[t_i, \hat{f}]$, she cannot improve her 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. □

7 IRM Is Not Necessary for Interim Rationalizable Implementation

In this section, we disprove the claim made in Oury and Tercieux (2012, footnote 4) that IRM is necessary for interim rationalizable implementation. We base our arguments on

the following example, which is built upon the example presented in Kunimoto and Saran (2020).

Example 7.1. 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 $T^* = T$ since for every state, there is a type of a player who puts positive probability on that state. Thus, we need not discuss equivalent SCFs in this setup.

There are six pure alternatives: $A = \{a, b, c, d, z, z^{'}\}$. The following tables list the payoffs of the two players:

a	t_2	$t_2^{'}$
t_1	$4, 4$	$4, 0$
$t_1^{'}$	$0, 0$	$4, 1$
$t_1^{''}$	$1, 1$	$4, 0$

b	t_2	$t_2^{'}$
t_1	$0, 0$	$3, 3$
$t_1^{'}$	$1, 1$	$2, 0$
$t_1^{''}$	$0, 0$	$2, 1$

c	t_2	$t_2^{'}$
t_1	$0, 0$	$3, 1$
$t_1^{'}$	$3, 3$	$3, 0$
$t_1^{''}$	$3, 3$	$3, 0$

d	t_2	$t_2^{'}$
t_1	$3, 4$	$2, 0$
$t_1^{'}$	$0, 3$	$3, 3$
$t_1^{''}$	$0, 3$	$3, 3$

z	t_2	$t_2^{'}$
t_1	$4, 1$	$2, 0$
$t_1^{'}$	$2, 2$	$5, 0$
$t_1^{''}$	$2, 2$	$2, 0$

$z^{'}$	t_2	$t_2^{'}$
t_1	$4, 0$	$4, 1$
$t_1^{'}$	$2, 0$	$2, 2$
$t_1^{''}$	$2, 0$	$5, 0$

The SCF f selects the alternative which maximizes the aggregate payoff in each state.

f	t_2	$t_2^{'}$
t_1	a	b
$t_1^{'}$	c	d
$t_1^{''}$	c	d

We first show that f fails BM.

Claim 7.2. *The SCF f violates BM.*

Proof. Consider the single-valued deception β^s such that

$$\beta_1^s(t_1) = t_1^{'}, \quad \beta_1^s(t_1^{'}) = t_1^{'}, \quad \beta_1^s(t_1^{''}) = t_1^{''},$$

and

$$\beta_2^s(t_2) = t_2^{'}, \quad \beta_2^s(t_2^{'}) = t_2^{'}.$$

First, consider player 2 of type t_2. There exists no $y : T_1 \to \Delta(A)$ such that

$$\begin{aligned} U_2(y \circ \beta_1^s | t_2) &= \frac{1}{3}u_2\big(y(t_1^{'}), (t_1, t_2)\big) + \frac{1}{3}u_2\big(y(t_1^{'}), (t_1^{'}, t_2)\big) + \frac{1}{3}u_2\big(y(t_1^{''}), (t_1^{''}, t_2)\big) \\ &> U_2(f \circ \beta^s | t_2) = \frac{1}{3}u_2\big(f(t_1^{'}, t_2^{'}), (t_1, t_2)\big) + \frac{1}{3}u_2\big(f(t_1^{'}, t_2^{'}), (t_1^{'}, t_2)\big) + \frac{1}{3}u_2\big(f(t_1^{''}, t_2^{'}), (t_1^{''}, t_2)\big), \end{aligned}$$

because $f(t_1^{'}, t_2^{'}) = f(t_1^{''}, t_2^{'}) = d$ is one of the best alternatives for player 2 of type t_2 in each state.

Second, consider player 2 of type $t_2^{'}$. There exists no $y : T_1 \to \Delta(A)$ such that

$$U_2(y \circ \beta_1^s | t_2^{'}) > U_2(f \circ \beta^s | t_2^{'}) \text{ and } U_2(f | t_2^{'}) \geq U_2(y | t_2^{'}).$$

Since $U_2(f \circ \beta^s | t_2^{'}) = u_2(f(t_1^{'}, t_2^{'}), (t_1^{'}, t_2^{'})) = U_2(f | t_2^{'})$, if the above inequalities were true, then we must have $U_2(y \circ \beta_1^s | t_2^{'}) > U_2(y | t_2^{'})$. But that is impossible because $U_2(y \circ \beta_1^s | t_2^{'}) = u_2\big(y(t_1^{'}), (t_1^{'}, t_2^{'})\big) = U_2(y | t_2^{'})$.

Third, consider player 1 of type t_1. Pick any $y : T_2 \to \Delta(A)$ such that

$$U_1(f | t_1) \geq U_1(y | t_1), \quad U_1(f | t_1^{'}) \geq U_1(y | t_1^{'}), \quad \text{and } U_1(f | t_1^{''}) \geq U_1(y | t_1^{''}).$$

The last two inequalities imply that

$$\begin{aligned} u_1\big(f(t_1^{'}, t_2^{'}), (t_1^{'}, t_2^{'})\big) &\geq u_1\big(y(t_2^{'}), (t_1^{'}, t_2^{'})\big) \\ u_1\big(f(t_1^{''}, t_2^{'}), (t_1^{''}, t_2^{'})\big) &\geq u_1\big(y(t_2^{'}), (t_1^{''}, t_2^{'})\big). \end{aligned}$$

These two inequalities lead to

$$2y(t_2^{'})[z] + y(t_2^{'})[a] \leq y(t_2^{'})[z^{'}] + y(t_2^{'})[b] \quad \text{and} \quad 2y(t_2^{'})[z^{'}] + y(t_2^{'})[a] \leq y(t_2^{'})[z] + y(t_2^{'})[b],$$

where $y(t_2^{'})[x]$ is the probability of alternative x in the lottery $y(t_2^{'})$. Summing these two inequalities, we obtain $y(t_2^{'})[z] + y(t_2^{'})[z^{'}] + 2y(t_2^{'})[a] \leq 2y(t_2^{'})[b]$.

In order to find the required preference reversal for type t_1, we must satisfy $U_1(y \circ \beta_2^s | t_1) > U_1(f \circ \beta^s | t_1)$, that is,

$$0.99u_1\big(y(t_2^{'}), (t_1, t_2)\big) + 0.01u_1\big(y(t_2^{'}), (t_1, t_2^{'})\big) > 0.99u_1\big(f(t_1^{'}, t_2^{'}), (t_1, t_2)\big) + 0.01u_1\big(f(t_1^{'}, t_2^{'}), (t_1, t_2^{'})\big).$$

The above inequality is translated into

$$0.99(y(t_2')[z] + y(t_2')[z'] + y(t_2')[a]) + 0.01(2y(t_2')[a] + y(t_2')[b] + y(t_2')[c] + 2y(t_2')[z']) \\ > 0.99(3y(t_2')[b] + 3y(t_2')[c]).$$

As $y(t_2')[z] + y(t_2')[z'] + 2y(t_2')[a] \leq 2y(t_2')[b]$, we must have $y(t_2')[z'] \leq 2y(t_2')[b]$. Plugging this into the left-hand side of the above inequality gives us

$$0.99(y(t_2')[z] + y(t_2')[z'] + y(t_2')[a]) + 0.01(2y(t_2')[a] + 5y(t_2')[b] + y(t_2')[c]) \\ > 0.99(3y(t_2')[b] + 3y(t_2')[c]).$$

We claim that this inequality is impossible to be satisfied. Now plugging $y(t_2')[z] + y(t_2')[z'] + 2y(t_2')[a] \leq 2y(t_2')[b]$ into the right-hand side of the above inequality, we obtain

$$-0.99y(t_2')[a] + 0.01(2y(t_2')[a] + 5y(t_2')[b] + y(t_2')[c]) > 0.99(y(t_2')[b] + 3y(t_2')[c]) \\ \Rightarrow -0.97y(t_2')[a] - 0.94y(t_2')[b] - 2.96y(t_2')[c] > 0,$$

which is indeed impossible.

Fourth, consider player 1 of type t_1'. There does not exist any $y : T_2 \to \Delta(A)$ such that

$$U_1(y \circ \beta_2^s | t_1') > U_1(f \circ \beta^s | t_1') \text{ and } U_1(f|t_1') \geq U_1(y|t_1').$$

Since $U_1(f \circ \beta^s | t_1') = u_1\big(f(t_1', t_2'), (t_1', t_2')\big) = U_1(f|t_1')$, if the above inequalities were true, then we must have $U_1(y \circ \beta_2^s | t_1') > U_1(y|t_1')$. But that is impossible because $U_1(y \circ \beta_2^s | t_1') = u_1\big(y(t_2'), (t_1', t_2')\big) = U_1(y|t_1')$.

Finally, consider player 1 of type t_1''. There does not exist any $y : T_2 \to \Delta(A)$ such that

$$U_1(y \circ \beta_2^s | t_1'') > U_1(f \circ \beta^s | t_1'') \text{ and } U_1(f|t_1'') \geq U_1(y|t_1'').$$

Since $U_1(f \circ \beta^s | t_1'') = u_1\big(f(t_1'', t_2'), (t_1'', t_2')\big) = U_1(f|t_1'')$, if the above inequalities were true, then we must have $U_1(y \circ \beta_2^s | t_1'') > U_1(y|t_1'')$. But that is impossible because $U_1(y \circ \beta_2^s | t_1'') = u_1\big(y(t_2'), (t_1'', t_2')\big) = U_1(y|t_1'')$.

We therefore conclude that the SCF f does not satisfy BM. □

Since we know from Lemma 5.8 that IRM implies BM, we state the following result without proof.

Claim 7.3. *The SCF f violates IRM.*

Next, we argue that f satisfies weak IRM.

Claim 7.4. *The SCF f satisfies weak IRM.*

Proof. First, we consider any unacceptable deception β such that either $t_1^{'} \in \beta_1(t_1)$ or $t_1^{''} \in \beta_1(t_1)$. Pick any belief $\psi_1 \in \Delta(T_2 \times T)$. Then

$$\begin{aligned}
& \sum_{\hat{t}_2,\tilde{t}} \psi_1(\hat{t}_2,\tilde{t}) u_1\big(f(t_1^{'},\tilde{t}_2),(t_1,\hat{t}_2)\big) \\
= & \sum_{\tilde{t}_1} \psi_1(t_2,\tilde{t}_1,t_2) u_1\big(f(t_1^{'},t_2),(t_1,t_2)\big) + \sum_{\tilde{t}_1} \psi_1(t_2,\tilde{t}_1,t_2^{'}) u_1\big(f(t_1^{'},t_2^{'}),(t_1,t_2)\big) \\
& + \sum_{\tilde{t}_1} \psi_1(t_2^{'},\tilde{t}_1,t_2) u_1\big(f(t_1^{'},t_2),(t_1,t_2^{'})\big) + \sum_{\tilde{t}_1} \psi_1(t_2^{'},\tilde{t}_1,t_2^{'}) u_1\big(f(t_1^{'},t_2^{'}),(t_1,t_2^{'})\big) \\
= & 3\sum_{\tilde{t}_1} \psi_1(t_2^{'},\tilde{t}_1,t_2) + 3\sum_{\tilde{t}_1} \psi_1(t_2,\tilde{t}_1,t_2^{'}) + 2\sum_{\tilde{t}_1} \psi_1(t_2^{'},\tilde{t}_1,t_2^{'}).
\end{aligned}$$

Since $f(t_1^{'},t_2) = f(t_1^{''},t_2) = c$ and $f(t_1^{'},t_2^{'}) = f(t_1^{''},t_2^{'}) = d$, we also obtain

$$\sum_{\hat{t}_2,\tilde{t}} \psi_1(\hat{t}_2,\tilde{t}) u_1\big(f(t_1^{''},\tilde{t}_2),(t_1,\hat{t}_2)\big) = 3\sum_{\tilde{t}_1} \psi_1(t_2^{'},\tilde{t}_1,t_2) + 3\sum_{\tilde{t}_1} \psi_1(t_2,\tilde{t}_1,t_2^{'}) + 2\sum_{\tilde{t}_1} \psi_1(t_2^{'},\tilde{t}_1,t_2^{'}).$$

Consider the SCF $f^{'}$ defined as follows:

$f^{'}$	t_2	$t_2^{'}$
t_1	a	$\frac{2}{3}z + \frac{1}{3}z^{'}$
$t_1^{'}$	a	$z^{'}$
$t_1^{''}$	a	$\frac{1}{5}c + \frac{4}{5}z$

It is straightforward to confirm that $f^{'}(\tilde{t}_1,\cdot) \in Y_1[\tilde{t}_1,f]$ for all $\tilde{t}_1 \in T_1$. Moreover,

$$\begin{aligned}
\sum_{\hat{t}_2,\tilde{t}} \psi_1(\hat{t}_2,\tilde{t}) u_1\big(f^{'}(\tilde{t}),(t_1,\hat{t}_2)\big) = & \sum_{\hat{t}_2,\tilde{t}_2} \psi_1(\hat{t}_2,t_1,\tilde{t}_2) u_1\big(f^{'}(t_1,\tilde{t}_2),(t_1,\hat{t}_2)\big) \\
& + \sum_{\hat{t}_2,\tilde{t}_2} \psi_1(\hat{t}_2,t_1^{'},\tilde{t}_2) u_1\big(f^{'}(t_1^{'},\tilde{t}_2),(t_1,\hat{t}_2)\big) \\
& + \sum_{\hat{t}_2,\tilde{t}_2} \psi_1(\hat{t}_2,t_1^{''},\tilde{t}_2) u_1\big(f^{'}(t_1^{''},\tilde{t}_2),(t_1,\hat{t}_2)\big).
\end{aligned}$$

We consider each term on the right-hand side of the above equation separately. The first term is:

$$\begin{aligned}
& \sum_{\hat{t}_2,\tilde{t}_2} \psi_1(\hat{t}_2,t_1,\tilde{t}_2) u_1\big(f^{'}(t_1,\tilde{t}_2),(t_1,\hat{t}_2)\big) \\
= & \quad \psi_1(t_2,t_1,t_2) u_1\big(f^{'}(t_1,t_2),(t_1,t_2)\big) + \psi_1(t_2,t_1,t_2^{'}) u_1\big(f^{'}(t_1,t_2^{'}),(t_1,t_2)\big)
\end{aligned}$$

$$
\begin{aligned}
& +\psi_1(t_2^{'}, t_1, t_2) u_1\big(f^{'}(t_1, t_2), (t_1, t_2^{'})\big) + \psi_1(t_2^{'}, t_1, t_2^{'}) u_1\big(f^{'}(t_1, t_2^{'}), (t_1, t_2^{'})\big) \\
= \quad & 4\psi_1(t_2, t_1, t_2) + 4\psi_1(t_2^{'}, t_1, t_2) + 4\psi_1(t_2, t_1, t_2^{'}) + \frac{8}{3}\psi_1(t_2^{'}, t_1, t_2^{'}).
\end{aligned}
$$

The second term is:

$$
\begin{aligned}
& \sum_{\hat{t}_2, \tilde{t}_2} \psi_1(\hat{t}_2, t_1^{'}, \tilde{t}_2) u_1\big(f^{'}(t_1^{'}, \tilde{t}_2), (t_1, \hat{t}_2)\big) \\
= \quad & \psi_1(t_2, t_1^{'}, t_2) u_1\big(f^{'}(t_1^{'}, t_2), (t_1, t_2)\big) + \psi_1(t_2, t_1^{'}, t_2^{'}) u_1\big(f^{'}(t_1^{'}, t_2^{'}), (t_1, t_2)\big) \\
& +\psi_1(t_2^{'}, t_1^{'}, t_2) u_1\big(f^{'}(t_1^{'}, t_2), (t_1, t_2^{'})\big) + \psi_1(t_2^{'}, t_1^{'}, t_2^{'}) u_1\big(f^{'}(t_1^{'}, t_2^{'}), (t_1, t_2^{'})\big) \\
= \quad & 4\psi_1(t_2, t_1^{'}, t_2) + 4\psi_1(t_2^{'}, t_1^{'}, t_2) + 4\psi_1(t_2, t_1^{'}, t_2^{'}) + 4\psi_1(t_2^{'}, t_1^{'}, t_2^{'}).
\end{aligned}
$$

The third term is:

$$
\begin{aligned}
& \sum_{\hat{t}_2, \tilde{t}_2} \psi_1(\hat{t}_2, t_1^{''}, \tilde{t}_2) u_1\big(f^{'}(t_1^{''}, \tilde{t}_2), (t_1, \hat{t}_2)\big) \\
= \quad & \psi_1(t_2, t_1^{''}, t_2) u_1\big(f^{'}(t_1^{''}, t_2), (t_1, t_2)\big) + \psi_1(t_2, t_1^{''}, t_2^{'}) u_1\big(f^{'}(t_1^{''}, t_2^{'}), (t_1, t_2)\big) \\
& +\psi_1(t_2^{'}, t_1^{''}, t_2) u_1\big(f^{'}(t_1^{''}, t_2), (t_1, t_2^{'})\big) + \psi_1(t_2^{'}, t_1^{''}, t_2^{'}) u_1\big(f^{'}(t_1^{''}, t_2^{'}), (t_1, t_2^{'})\big) \\
= \quad & 4\psi_1(t_2, t_1^{''}, t_2) + 4\psi_1(t_2^{'}, t_1^{''}, t_2) + \frac{16}{5}\psi_1(t_2, t_1^{''}, t_2^{'}) + \frac{11}{5}\psi_1(t_2^{'}, t_1^{''}, t_2^{'}).
\end{aligned}
$$

Summing the three terms, we get

$$
\begin{aligned}
& \sum_{\hat{t}_2, \tilde{t}} \psi_1(\hat{t}_2, \tilde{t}) u_1\big(f^{'}(\tilde{t}), (t_1, \hat{t}_2)\big) \\
= \quad & 4\sum_{\tilde{t}_1} \psi_1(t_2, \tilde{t}_1, t_2) + 4\sum_{\tilde{t}_1} \psi_1(t_2^{'}, \tilde{t}_1, t_2) \\
& + \left(4\psi_1(t_2, t_1, t_2^{'}) + 4\psi_1(t_2, t_1^{'}, t_2^{'}) + \frac{16}{5}\psi_1(t_2, t_1^{''}, t_2^{'})\right) \\
& + \left(\frac{8}{3}\psi_1(t_2^{'}, t_1, t_2^{'}) + 4\psi_1(t_2^{'}, t_1^{'}, t_2^{'}) + \frac{11}{5}\psi_1(t_2^{'}, t_1^{''}, t_2^{'})\right) \\
\geq \quad & 4\sum_{\tilde{t}_1} \psi_1(t_2, \tilde{t}_1, t_2) + 4\sum_{\tilde{t}_1} \psi_1(t_2^{'}, \tilde{t}_1, t_2) + \frac{16}{5}\sum_{\tilde{t}_1} \psi_1(t_2, \tilde{t}_1, t_2^{'}) + \frac{11}{5}\sum_{\tilde{t}_1} \psi_1(t_2^{'}, \tilde{t}_1, t_2^{'}) \\
> \quad & 3\sum_{\tilde{t}_1} \psi_1(t_2^{'}, \tilde{t}_1, t_2) + 3\sum_{\tilde{t}_1} \psi_1(t_2, \tilde{t}_1, t_2^{'}) + 2\sum_{\tilde{t}_1} \psi_1(t_2^{'}, \tilde{t}_1, t_2^{'}).
\end{aligned}
$$

We therefore conclude that

$$
\sum_{\hat{t}_2, \tilde{t}} \psi_1(\hat{t}_2, \tilde{t}) u_1\big(f^{'}(\tilde{t}), (t_1, \hat{t}_2)\big) > \sum_{\hat{t}_2, \tilde{t}} \psi_1(\hat{t}_2, \tilde{t}) u_1\big(f(t_1^{'}, \tilde{t}_2), (t_1, \hat{t}_2)\big).
$$

Similarly,

$$\sum_{\hat{t}_2,\tilde{t}} \psi_1(\hat{t}_2,\tilde{t}) u_1\big(f^{'}(\tilde{t}),(t_1,\hat{t}_2)\big) > \sum_{\hat{t}_2,\tilde{t}} \psi_1(\hat{t}_2,\tilde{t}) u_1\big(f(t_1^{''},\tilde{t}_2),(t_1,\hat{t}_2)\big).$$

It follows that any unacceptable deception β satisfying $t_1^{'} \in \beta_1(t_1)$ is weakly refutable using the tuple $(1,t_1,t_1^{'})$ whereas any unacceptable deception β satisfying $t_1^{''} \in \beta_1(t_1)$ is weakly refutable using the tuple $(1,t_1,t_1^{''})$.

Second, we consider any unacceptable deception β such that $t_2^{'} \in \beta_2(t_2)$ and $\beta_1(t_1) = \{t_1\}$. Pick any belief $\psi_2 \in \Delta(T_1 \times T)$ such that $\psi_2(\hat{t}_1,\tilde{t}) > 0 \Rightarrow \tilde{t}_1 \in \beta_1(\hat{t}_1)$ and $\pi_2(t_2)[\hat{t}_1] = \sum_{\tilde{t}} \psi_2(\hat{t}_1,\tilde{t})$ for all $\hat{t}_1$. Then we have $\psi_2(t_1,\tilde{t}) = 0$ whenever $\tilde{t}_1 \neq t_1$ and $\sum_{\tilde{t}_2} \psi_2(t_1,t_1,\tilde{t}_2) = 1/3$. Therefore,

$$\begin{aligned}
& \sum_{\hat{t}_1,\tilde{t}} \psi_2(\hat{t}_1,\tilde{t}) u_2\big(f(\tilde{t}_1,t_2^{'}),(\hat{t}_1,t_2)\big) \\
= & \sum_{\tilde{t}} \psi_2(t_1,\tilde{t}) u_2\big(f(\tilde{t}_1,t_2^{'}),(t_1,t_2)\big) + \sum_{\tilde{t}} \psi_2(t_1^{'},\tilde{t}) u_2\big(f(\tilde{t}_1,t_2^{'}),(t_1^{'},t_2)\big) \\
& + \sum_{\tilde{t}} \psi_2(t_1^{''},\tilde{t}) u_2\big(f(\tilde{t}_1,t_2^{'}),(t_1^{''},t_2)\big) \\
= & \sum_{\tilde{t}_2} \psi_2(t_1,t_1,\tilde{t}_2) u_2\big(f(t_1,t_2^{'}),(t_1,t_2)\big) + \sum_{\tilde{t}_2} \psi_2(t_1^{'},t_1,\tilde{t}_2) u_2\big(f(t_1,t_2^{'}),(t_1^{'},t_2)\big) \\
& + \sum_{\tilde{t}_2} \psi_2(t_1^{'},t_1^{'},\tilde{t}_2) u_2\big(f(t_1^{'},t_2^{'}),(t_1^{'},t_2)\big) + \sum_{\tilde{t}_2} \psi_2(t_1^{'},t_1^{''},\tilde{t}_2) u_2\big(f(t_1^{''},t_2^{'}),(t_1^{'},t_2)\big) \\
& + \sum_{\tilde{t}_2} \psi_2(t_1^{''},t_1,\tilde{t}_2) u_2\big(f(t_1,t_2^{'}),(t_1^{''},t_2)\big) + \sum_{\tilde{t}_2} \psi_2(t_1^{''},t_1^{'},\tilde{t}_2) u_2\big(f(t_1^{'},t_2^{'}),(t_1^{''},t_2)\big) \\
& + \sum_{\tilde{t}_2} \psi_2(t_1^{''},t_1^{''},\tilde{t}_2) u_2\big(f(t_1^{''},t_2^{'}),(t_1^{''},t_2)\big) \\
= & \sum_{\tilde{t}_2} \psi_2(t_1^{'},t_1,\tilde{t}_2) + 3\sum_{\tilde{t}_2} \big(\psi_2(t_1^{'},t_1^{'},\tilde{t}_2) + \psi_2(t_1^{'},t_1^{''},\tilde{t}_2) + \psi_2(t_1^{''},t_1^{'},\tilde{t}_2) + \psi_2(t_1^{''},t_1^{''},\tilde{t}_2)\big).
\end{aligned}$$

Consider the SCF $f^{'}$ defined as follows:

$f^{'}$	t_2	$t_2^{'}$
t_1	z	z
$t_1^{'}$	d	d
$t_1^{''}$	d	d

It is straightforward to confirm that $f^{'}(\cdot,\tilde{t}_2) \in Y_2[\tilde{t}_2,f]$ for all $\tilde{t}_2 \in T_2$. Moreover, because

$f'(\tilde{t}_1, t_2) = f'(\tilde{t}_1, t_2')$ for all $\tilde{t}_1 \in T_1$, we have

$$\sum_{\hat{t}_1,\tilde{t}} \psi_2(\hat{t}_1,\tilde{t}) u_2\big(f'(\tilde{t}), (\hat{t}_1, t_2)\big) = \sum_{\hat{t}_1,\tilde{t}} \psi_2(\hat{t}_1,\tilde{t}) u_2\big(f'(\tilde{t}_1, t_2'), (\hat{t}_1, t_2)\big).$$

Since $\psi_2(t_1,\tilde{t}) = 0$ whenever $\tilde{t}_1 \neq t_1$ and $\sum_{\tilde{t}_2} \psi_2(t_1, t_1, \tilde{t}_2) = 1/3$, by applying here similar arguments as in the case of the SCF f, we obtain that

$$\begin{aligned}
&\sum_{\hat{t}_1,\tilde{t}} \psi_2(\hat{t}_1,\tilde{t}) u_2\big(f'(\tilde{t}_1, t_2'), (\hat{t}_1, t_2)\big) \\
= \;& \sum_{\tilde{t}_2} \psi_2(t_1, t_1, \tilde{t}_2) + 2\sum_{\tilde{t}_2} \big(\psi_2(t_1', t_1, \tilde{t}_2) + \psi_2(t_1'', t_1, \tilde{t}_2)\big) \\
& +3\sum_{\tilde{t}_2} \big(\psi_2(t_1', t_1', \tilde{t}_2) + \psi_2(t_1', t_1'', \tilde{t}_2) + \psi_2(t_1'', t_1', \tilde{t}_2) + \psi_2(t_1'', t_1'', \tilde{t}_2)\big) \\
> \;& \sum_{\tilde{t}_2} \psi_2(t_1', t_1, \tilde{t}_2) + 3\sum_{\tilde{t}_2} \big(\psi_2(t_1', t_1', \tilde{t}_2) + \psi_2(t_1', t_1'', \tilde{t}_2) + \psi_2(t_1'', t_1', \tilde{t}_2) + \psi_2(t_1'', t_1'', \tilde{t}_2)\big).
\end{aligned}$$

It follows that

$$\sum_{\hat{t}_1,\tilde{t}} \psi_2(\hat{t}_1,\tilde{t}) u_2\big(f'(\tilde{t}), (\hat{t}_1, t_2)\big) > \sum_{\hat{t}_1,\tilde{t}} \psi_2(\hat{t}_1,\tilde{t}) u_2\big(f(\tilde{t}_1, t_2'), (\hat{t}_1, t_2)\big).$$

Therefore, any unacceptable deception β such that $t_2' \in \beta_2(t_2)$ and $\beta_1(t_1) = \{t_1\}$ is weakly refutable using the tuple $(2, t_2, t_2')$.

Third, we consider any unacceptable deception β such that $t_2 \in \beta_2(t_2')$ and $\beta_1(t_1) = \{t_1\}$. Pick any belief $\psi_2 \in \Delta(T_1 \times T)$ such that $\psi_2(\hat{t}_1,\tilde{t}) > 0 \Rightarrow \tilde{t}_1 \in \beta_1(\hat{t}_1)$. Then we have that $\psi_2(t_1,\tilde{t}) = 0$ whenever $\tilde{t}_1 \neq t_1$. Therefore,

$$\begin{aligned}
&\sum_{\hat{t}_1,\tilde{t}} \psi_2(\hat{t}_1,\tilde{t}) u_2\big(f(\tilde{t}_1, t_2), (\hat{t}_1, t_2')\big) \\
= \;& \sum_{\tilde{t}} \psi_2(t_1,\tilde{t}) u_2\big(f(\tilde{t}_1, t_2), (t_1, t_2')\big) + \sum_{\tilde{t}} \psi_2(t_1',\tilde{t}) u_2\big(f(\tilde{t}_1, t_2), (t_1', t_2')\big) \\
& + \sum_{\tilde{t}} \psi_2(t_1'',\tilde{t}) u_2\big(f(\tilde{t}_1, t_2), (t_1'', t_2')\big) \\
= \;& \sum_{\tilde{t}_2} \psi_2(t_1, t_1, \tilde{t}_2) u_2\big(f(t_1, t_2), (t_1, t_2')\big) + \sum_{\tilde{t}_2} \psi_2(t_1', t_1, \tilde{t}_2) u_2\big(f(t_1, t_2), (t_1', t_2')\big) \\
& + \sum_{\tilde{t}_2} \psi_2(t_1', t_1', \tilde{t}_2) u_2\big(f(t_1', t_2), (t_1', t_2')\big) + \sum_{\tilde{t}_2} \psi_2(t_1', t_1'', \tilde{t}_2) u_2\big(f(t_1'', t_2), (t_1', t_2')\big) \\
& + \sum_{\tilde{t}_2} \psi_2(t_1'', t_1, \tilde{t}_2) u_2\big(f(t_1, t_2), (t_1'', t_2')\big) + \sum_{\tilde{t}_2} \psi_2(t_1'', t_1', \tilde{t}_2) u_2\big(f(t_1', t_2), (t_1'', t_2')\big)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{\tilde{t}_2}\psi_2(t_1^{''},t_1^{''},\tilde{t}_2)u_2\big(f(t_1^{''},t_2),(t_1^{''},t_2^{'})\big)\\
=\ &\sum_{\tilde{t}_2}\psi_2(t_1^{'},t_1,\tilde{t}_2).
\end{aligned}
$$

Consider the constant SCF $f^{'}$ such that $f^{'}(\tilde{t}) = \frac{1}{4}b + \frac{3}{4}z^{'}$ for all $\tilde{t} \in T$. It is straightforward to confirm that $f^{'}(\cdot,\tilde{t}_2) \in Y_2[\tilde{t}_2, f]$ for all $\tilde{t}_2 \in T_2$. Moreover, because $f^{'}(\tilde{t}_1,t_2) = f^{'}(\tilde{t}_1,t_2^{'})$ for all $\tilde{t}_1 \in T_1$, we have

$$
\sum_{\hat{t}_1,\tilde{t}}\psi_2(\hat{t}_1,\tilde{t})u_2\big(f^{'}(\tilde{t}),(\hat{t}_1,t_2^{'})\big) = \sum_{\hat{t}_1,\tilde{t}}\psi_2(\hat{t}_1,\tilde{t})u_2\big(f^{'}(\tilde{t}_1,t_2),(\hat{t}_1,t_2^{'})\big).
$$

Since $\psi_2(t_1,\tilde{t}) = 0$ whenever $\tilde{t}_1 \neq t_1$, by applying here similar arguments as in the case of the SCF f, we obtain that

$$
\begin{aligned}
&\sum_{\hat{t}_1,\tilde{t}}\psi_2(\hat{t}_1,\tilde{t})u_2\big(f^{'}(\tilde{t}_1,t_2),(\hat{t}_1,t_2^{'})\big)\\
=\ &\frac{3}{2}\sum_{\tilde{t}_2}\psi_2(t_1,t_1,\tilde{t}_2)+\frac{3}{2}\sum_{\tilde{t}_2}\big(\psi_2(t_1^{'},t_1,\tilde{t}_2)+\psi_2(t_1^{'},t_1^{'},\tilde{t}_2)+\psi_2(t_1^{'},t_1^{''},\tilde{t}_2)\big)\\
&+\frac{1}{4}\sum_{\tilde{t}_2}\big(\psi_2(t_1^{''},t_1,\tilde{t}_2)+\psi_2(t_1^{''},t_1^{'},\tilde{t}_2)+\psi_2(t_1^{''},t_1^{''},\tilde{t}_2)\big)\\
>\ &\sum_{\tilde{t}_2}\psi_2(t_1^{'},t_1,\tilde{t}_2).
\end{aligned}
$$

It follows that

$$
\sum_{\hat{t}_1,\tilde{t}}\psi_2(\hat{t}_1,\tilde{t})u_2\big(f^{'}(\tilde{t}),(\hat{t}_1,t_2^{'})\big) > \sum_{\hat{t}_1,\tilde{t}}\psi_2(\hat{t}_1,\tilde{t})u_2\big(f(\tilde{t}_1,t_2),(\hat{t}_1,t_2^{'})\big).
$$

Therefore, any unacceptable deception β such that $t_2 \in \beta_2(t_2^{'})$ and $\beta_1(t_1) = \{t_1\}$ is weakly refutable using the tuple $(2,t_2^{'},t_2)$.

Fourth, we consider any unacceptable deception such that $\beta_1(t_1) = \{t_1\}$, $\beta_2(t_2) = \{t_2\}$, and $\beta_2(t_2^{'}) = \{t_2^{'}\}$. Such a deception involves either $t_1 \in \beta_1(t_1^{'})$ or $t_1 \in \beta_1(t_1^{''})$. Then the fact that f satisfies SIRBIC implies that β is weakly refutable. We show this formally for the case when $t_1 \in \beta_1(t_1^{'})$ and we skip the case when $t_1 \in \beta_1(t_1^{''})$, as we can show it similarly. So suppose $t_1 \in \beta_1(t_1^{'})$. Pick any belief $\psi_1 \in \Delta(T_2 \times T)$ such that $\psi_1(\hat{t}_2,\tilde{t}) >$

$0 \Rightarrow \tilde{t}_2 \in \beta_2(\hat{t}_2)$. Then we have that $\psi_1(\hat{t}_2, \tilde{t}) = 0$ whenever $\tilde{t}_2 \neq \hat{t}_2$. Therefore,

$$\begin{aligned}
& \sum_{\hat{t}_2, \tilde{t}} \psi_1(\hat{t}_2, \tilde{t}) u_1\big(f(t_1, \tilde{t}_2), (t_1', \hat{t}_2)\big) \\
= & \sum_{\tilde{t}_1} \psi_1(t_2, \tilde{t}_1, t_2) u_1(f(t_1, t_2), (t_1', t_2)) + \sum_{\tilde{t}_1} \psi_1(t_2', \tilde{t}_1, t_2') u_1(f(t_1, t_2'), (t_1', t_2')) \\
= & \; 2 \sum_{\tilde{t}_1} \psi_1(t_2', \tilde{t}_1, t_2').
\end{aligned}$$

Consider the SCF f' defined as follows:

f'	t_2	t_2'
t_1	c	d
t_1'	c	d
t_1''	c	d

It is straightforward to confirm that $f'(\tilde{t}_1, \cdot) \in Y_1[\tilde{t}_1, f]$ for all $\tilde{t}_1 \in T_1$. Moreover, since $\psi_1(\hat{t}_2, \tilde{t}) = 0$ whenever $\tilde{t}_2 \neq \hat{t}_2$,

$$\begin{aligned}
& \sum_{\hat{t}_2, \tilde{t}} \psi_1(\hat{t}_2, \tilde{t}) u_1\big(f'(\tilde{t}), (t_1', \hat{t}_2)\big) \\
= & \sum_{\tilde{t}_1} \psi_1(t_2, \tilde{t}_1, t_2) u_1(f'(\tilde{t}_1, t_2), (t_1', t_2)) + \sum_{\tilde{t}_1} \psi_1(t_2', \tilde{t}_1, t_2') u_1(f'(\tilde{t}_1, t_2'), (t_1', t_2') \\
= & \; 3 \sum_{\tilde{t}_1} \big(\psi_1(t_2, \tilde{t}_1, t_2) + \psi_1(t_2', \tilde{t}_1, t_2')\big) \\
> & \; 2 \sum_{\tilde{t}_1} \psi_1(t_2', \tilde{t}_1, t_2') \\
= & \sum_{\hat{t}_2, \tilde{t}} \psi_1(\hat{t}_2, \tilde{t}) u_1\big(f(t_1, \tilde{t}_2), (t_1', \hat{t}_2)\big).
\end{aligned}$$

It follows that the deception β is weakly refutable using the tuple $(1, t_1', t_1)$.

We thus conclude that every unacceptable deception is weakly refutable, and hence f satisfies weak IRM. □

We now check that the SCF f satisfies weak NWR.

Claim 7.5. *The SCF f satisfies weak NWR.*

Proof. First, we consider player 1 of type t_1. Let $y : T_2 \to \Delta(A)$ be such that $y(t_2) = a$ and $y(t_2') = z$. Also, let $y' : T_2 \to \Delta(A)$ be such that $y'(t_2) = b$ and $y'(t_2') = d$. It is

straightforward to confirm that $y, y^{'} \in Y_1^w[t_1, f]$. Now,

$$\begin{aligned}
&\sum_{\tilde{t}_2, \tilde{t}_2^{'}} \phi_1(\tilde{t}_2, \tilde{t}_2^{'}) u_1\big(y(\tilde{t}_2^{'}), (t_1, \tilde{t}_2)\big) \\
= \;& \phi_1(t_2, t_2) u_1\big(y(t_2), (t_1, t_2)\big) + \phi_1(t_2, t_2^{'}) u_1\big(y(t_2^{'}), (t_1, t_2)\big) \\
& + \phi_1(t_2^{'}, t_2) u_1\big(y(t_2), (t_1, t_2^{'})\big) + \phi_1(t_2^{'}, t_2^{'}) u_1\big(y(t_2^{'}), (t_1, t_2^{'})\big) \\
= \;& 4\phi_1(t_2, t_2) + 4\phi_1(t_2, t_2^{'}) + 4\phi_1(t_2^{'}, t_2) + 2\phi_1(t_2^{'}, t_2^{'}).
\end{aligned}$$

whereas

$$\begin{aligned}
&\sum_{\tilde{t}_2, \tilde{t}_2^{'}} \phi_1(\tilde{t}_2, \tilde{t}_2^{'}) u_1\big(y^{'}(\tilde{t}_2^{'}), (t_1, \tilde{t}_2)\big) \\
= \;& \phi_1(t_2, t_2) u_1\big(y^{'}(t_2), (t_1, t_2)\big) + \phi_1(t_2, t_2^{'}) u_1\big(y^{'}(t_2^{'}), (t_1, t_2)\big) \\
& + \phi_1(t_2^{'}, t_2) u_1\big(y^{'}(t_2), (t_1, t_2^{'})\big) + \phi_1(t_2^{'}, t_2^{'}) u_1\big(y^{'}(t_2^{'}), (t_1, t_2^{'})\big) \\
= \;& 3\phi_1(t_2, t_2^{'}) + 3\phi_1(t_2^{'}, t_2) + 2\phi_1(t_2^{'}, t_2^{'}).
\end{aligned}$$

We therefore have that

$$\sum_{\tilde{t}_2, \tilde{t}_2^{'}} \phi_1(\tilde{t}_2, \tilde{t}_2^{'}) u_1\big(y(\tilde{t}_2^{'}), (t_1, \tilde{t}_2)\big) = \sum_{\tilde{t}_2, \tilde{t}_2^{'}} \phi_1(\tilde{t}_2, \tilde{t}_2^{'}) u_1\big(y^{'}(\tilde{t}_2^{'}), (t_1, \tilde{t}_2)\big) \Leftrightarrow \phi_1(t_2^{'}, t_2^{'}) = 1.$$

Thus, for all $\phi_1 \in \Delta(T_2 \times T_2)$ such that $\phi_1(t_2^{'}, t_2^{'}) < 1$, we have found $y, y^{'} \in Y_1^w[t_1]$ that satisfy the requirement for weak NWR. If ϕ_1 is such that $\phi_1(t_2^{'}, t_2^{'}) = 1$, then we define $y : T_2 \to \Delta(A)$ such that $y(t_2) = y(t_2^{'}) = b$ and $y^{'} : T_2 \to \Delta(A)$ such that $y^{'}(t_2) = y^{'}(t_2^{'}) = d$. It is straightforward to confirm that $y, y^{'} \in Y_1^w[t_1, f]$. Since $\phi_1(t_2^{'}, t_2^{'}) = 1$, $u_1(y(t_2^{'}), (t_1, t_2^{'})) = u_1(b, (t_1, t_2^{'})) = 3$ and $u_1(y^{'}(t_2^{'}), (t_1, t_2^{'})) = u_1(d, (t_1, t_2^{'})) = 2$, we obtain

$$\sum_{\tilde{t}_2, \tilde{t}_2^{'}} \phi_1(\tilde{t}_2, \tilde{t}_2^{'}) u_1\big(y(\tilde{t}_2^{'}), (t_1, \tilde{t}_2)\big) > \sum_{\tilde{t}_2, \tilde{t}_2^{'}} \phi_1(\tilde{t}_2, \tilde{t}_2^{'}) u_1\big(y^{'}(\tilde{t}_2^{'}), (t_1, \tilde{t}_2)\big).$$

Thus, if ϕ_1 is such that $\phi_1(t_2^{'}, t_2^{'}) = 1$, then the requirement for weak NWR is also satisfied.

Second, we consider player 1 of type $t_1^{'}$. Then we define $y : T_2 \to \Delta(A)$ such that $y(t_2) = y(t_2^{'}) = c$ and $y^{'} : T_2 \to \Delta(A)$ such that $y^{'}(t_2) = y^{'}(t_2^{'}) = b$. It is straightforward to confirm that $y, y^{'} \in Y_1^w[t_1^{'}, f]$. Fix $\phi_1 \in \Delta(T_2 \times T_2)$. Now,

$$\sum_{\tilde{t}_2, \tilde{t}_2^{'}} \phi_1(\tilde{t}_2, \tilde{t}_2^{'}) u_1\big(y(\tilde{t}_2^{'}), (t_1^{'}, \tilde{t}_2)\big)$$

$$
\begin{aligned}
&= \phi_1(t_2,t_2)u_1\big(y(t_2),(t_1^{'},t_2)\big)+\phi_1(t_2,t_2^{'})u_1\big(y(t_2^{'}),(t_1^{'},t_2)\big)\\
&\quad +\phi_1(t_2^{'},t_2)u_1\big(y(t_2),(t_1^{'},t_2^{'})\big)+\phi_1(t_2^{'},t_2^{'})u_1\big(y(t_2^{'}),(t_1^{'},t_2^{'})\big)\\
&= 3\phi_1(t_2,t_2)+3\phi_1(t_2,t_2^{'})+3\phi_1(t_2^{'},t_2)+3\phi_1(t_2^{'},t_2^{'})
\end{aligned}
$$

whereas

$$
\begin{aligned}
&\sum_{\tilde{t}_2,\tilde{t}_2^{'}}\phi_1(\tilde{t}_2,\tilde{t}_2^{'})u_1\big(y^{'}(\tilde{t}_2^{'}),(t_1^{'},\tilde{t}_2)\big)\\
&= \phi_1(t_2,t_2)u_1\big(y^{'}(t_2),(t_1^{'},t_2)\big)+\phi_1(t_2,t_2^{'})u_1\big(y^{'}(t_2^{'}),(t_1^{'},t_2)\big)\\
&\quad +\phi_1(t_2^{'},t_2)u_1\big(y^{'}(t_2),(t_1^{'},t_2^{'})\big)+\phi_1(t_2^{'},t_2^{'})u_1\big(y^{'}(t_2^{'}),(t_1^{'},t_2^{'})\big)\\
&= \phi_1(t_2,t_2)+\phi_1(t_2,t_2^{'})+2\phi_1(t_2^{'},t_2)+2\phi_1(t_2^{'},t_2^{'}).
\end{aligned}
$$

This implies that for any $\phi_1 \in \Delta(T_2 \times T_2)$,

$$
\sum_{\tilde{t}_2,\tilde{t}_2^{'}}\phi_1(\tilde{t}_2,\tilde{t}_2^{'})u_1\big(y(\tilde{t}_2^{'}),(t_1^{'},\tilde{t}_2)\big) > \sum_{\tilde{t}_2,\tilde{t}_2^{'}}\phi_1(\tilde{t}_2,\tilde{t}_2^{'})u_1\big(y^{'}(\tilde{t}_2^{'}),(t_1^{'},\tilde{t}_2)\big).
$$

Thus, the requirement for weak NWR is met.

Third, we consider player 1 of type $t_1^{''}$. Once again, we define $y : T_2 \to \Delta(A)$ such that $y(t_2) = y(t_2^{'}) = c$ and $y^{'} : T_2 \to \Delta(A)$ such that $y^{'}(t_2) = y^{'}(t_2^{'}) = b$. It is straightforward to confirm that $y, y^{'} \in Y_1^w[t_1^{''}, f]$. Fix $\phi_1 \in \Delta(T_2 \times T_2)$. Now

$$
\begin{aligned}
&\sum_{\tilde{t}_2,\tilde{t}_2^{'}}\phi_1(\tilde{t}_2,\tilde{t}_2^{'})u_1\big(y(\tilde{t}_2^{'}),(t_1^{''},\tilde{t}_2)\big)\\
&= \phi_1(t_2,t_2)u_1\big(y(t_2),(t_1^{''},t_2)\big)+\phi_1(t_2,t_2^{'})u_1\big(y(t_2^{'}),(t_1^{''},t_2)\big)\\
&\quad +\phi_1(t_2^{'},t_2)u_1\big(y(t_2),(t_1^{''},t_2^{'})\big)+\phi_1(t_2^{'},t_2^{'})u_1\big(y(t_2^{'}),(t_1^{''},t_2^{'})\big)\\
&= 3\phi_1(t_2,t_2)+3\phi_1(t_2,t_2^{'})+3\phi_1(t_2^{'},t_2)+3\phi_1(t_2^{'},t_2^{'})
\end{aligned}
$$

whereas

$$
\begin{aligned}
&\sum_{\tilde{t}_2,\tilde{t}_2^{'}}\phi_1(\tilde{t}_2,\tilde{t}_2^{'})u_1\big(y^{'}(\tilde{t}_2^{'}),(t_1^{''},\tilde{t}_2)\big)\\
&= \phi_1(t_2,t_2)u_1\big(y^{'}(t_2),(t_1^{''},t_2)\big)+\phi_1(t_2,t_2^{'})u_1\big(y^{'}(t_2^{'}),(t_1^{''},t_2)\big)\\
&\quad +\phi_1(t_2^{'},t_2)u_1\big(y^{'}(t_2),(t_1^{''},t_2^{'})\big)+\phi_1(t_2^{'},t_2^{'})u_1\big(y^{'}(t_2^{'}),(t_1^{''},t_2^{'})\big)\\
&= 2\phi_1(t_2^{'},t_2)+2\phi_1(t_2^{'},t_2^{'}).
\end{aligned}
$$

This implies that for any $\phi_1 \in \Delta(T_2 \times T_2)$,

$$\sum_{\tilde{t}_2, \tilde{t}_2^{'}} \phi_1(\tilde{t}_2, \tilde{t}_2^{'}) u_1\big(y(\tilde{t}_2^{'}), (t_1^{''}, \tilde{t}_2)\big) > \sum_{\tilde{t}_2, \tilde{t}_2^{'}} \phi_1(\tilde{t}_2, \tilde{t}_2^{'}) u_1\big(y^{'}(\tilde{t}_2^{'}), (t_1^{''}, \tilde{t}_2)\big).$$

Thus, the requirement for weak NWR is satisfied as well.

Fourth, we consider player 2 of type t_2. Then we define $y : T_1 \to \Delta(A)$ such that $y(t_1) = y(t_1^{'}) = y(t_1^{''}) = \frac{1}{2}a + \frac{1}{2}c$ and $y^{'} : T_1 \to \Delta(A)$ such that $y^{'}(t_1) = y^{'}(t_1^{'}) = y(t_1^{''}) = b$. It is straightforward to confirm that $y, y^{'} \in Y_2^w[t_2, f]$. Fix $\phi_2 \in \Delta(T_2 \times T_2)$. Now

$$\begin{aligned}
&\sum_{\tilde{t}_1, \tilde{t}_1^{'}} \phi_2(\tilde{t}_1, \tilde{t}_1^{'}) u_2(y(\tilde{t}_1^{'}), (\tilde{t}_1, t_2)) \\
= \;& \phi_2(t_1, t_1) u_2\big(y(t_1), (t_1, t_2)\big) + \phi_2(t_1, t_1^{'}) u_2\big(y(t_1^{'}), (t_1, t_2)\big) + \phi_2(t_1, t_1^{''}) u_2(y(t_1^{''}), (t_1, t_2)) \\
& + \phi_2(t_1^{'}, t_1) u_2\big(y(t_1), (t_1^{'}, t_2)\big) + \phi_2(t_1^{'}, t_1^{'}) u_2\big(y(t_1^{'}), (t_1^{'}, t_2)\big) + \phi_2(t_1^{'}, t_1^{''}) u_2(y(t_1^{''}), (t_1^{'}, t_2)) \\
& + \phi_2(t_1^{''}, t_1) u_2\big(y(t_1), (t_1^{''}, t_2)\big) + \phi_2(t_1^{''}, t_1^{'}) u_2\big(y(t_1^{'}), (t_1^{''}, t_2)\big) + \phi_2(t_1^{''}, t_1^{''}) u_2(y(t_1^{''}), (t_1^{''}, t_2)) \\
= \;& 2(\phi_1(t_1, t_1) + \phi_1(t_1, t_1^{'}) + \phi_1(t_1, t_1^{''})) + \frac{3}{2}(\phi_2(t_1^{'}, t_1) + \phi_2(t_1^{'}, t_1^{'}) + \phi_2(t_1^{'}, t_1^{''})) \\
& + 2(\phi_2(t_1^{''}, t_1) + \phi_2(t_1^{''}, t_1^{'}) + \phi_2(t_1^{''}, t_1^{''})),
\end{aligned}$$

whereas

$$\begin{aligned}
&\sum_{\tilde{t}_1, \tilde{t}_1^{'}} \phi_2(\tilde{t}_1, \tilde{t}_1^{'}) u_2(y^{'}(\tilde{t}_1^{'}), (\tilde{t}_1, t_2)) \\
= \;& \phi_2(t_1, t_1) u_2\big(y^{'}(t_1), (t_1, t_2)\big) + \phi_2(t_1, t_1^{'}) u_2\big(y^{'}(t_1^{'}), (t_1, t_2)\big) + \phi_2(t_1, t_1^{''}) u_2(y^{'}(t_1^{''}), (t_1, t_2)) \\
& + \phi_2(t_1^{'}, t_1) u_2\big(y^{'}(t_1), (t_1^{'}, t_2)\big) + \phi_2(t_1^{'}, t_1^{'}) u_2\big(y^{'}(t_1^{'}), (t_1^{'}, t_2)\big) + \phi_2(t_1^{'}, t_1^{''}) u_2(y^{'}(t_1^{''}), (t_1^{'}, t_2)) \\
& + \phi_2(t_1^{''}, t_1) u_2\big(y^{'}(t_1), (t_1^{''}, t_2)\big) + \phi_2(t_1^{''}, t_1^{'}) u_2\big(y^{'}(t_1^{'}), (t_1^{''}, t_2)\big) + \phi_2(t_1^{''}, t_1^{''}) u_2(y^{'}(t_1^{''}), (t_1^{''}, t_2)) \\
= \;& \phi_2(t_1^{'}, t_1) + \phi_2(t_1^{'}, t_1^{'}) + \phi_2(t_1^{'}, t_1^{''}).
\end{aligned}$$

This implies that for any $\phi_2 \in \Delta(T_1 \times T_1)$,

$$\sum_{\tilde{t}_1, \tilde{t}_1^{'}} \phi_2(\tilde{t}_1, \tilde{t}_1^{'}) u_2(y(\tilde{t}_1^{'}), (\tilde{t}_1, t_2)) > \sum_{\tilde{t}_1, \tilde{t}_1^{'}} \phi_2(\tilde{t}_1, \tilde{t}_1^{'}) u_2(y^{'}(\tilde{t}_1^{'}), (\tilde{t}_1, t_2)).$$

Thus, the requirement for weak NWR is also met.

And finally, we consider player 2 of type $t_2^{'}$. Then we define $y : T_1 \to \Delta(A)$ such that $y(t_1) = y(t_1^{'}) = y(t_1^{''}) = \frac{1}{2}b + \frac{1}{2}d$ and $y^{'} : T_1 \to \Delta(A)$ such that $y^{'}(t_1) = y^{'}(t_1^{'}) = y(t_1^{''}) = c$.

It is straightforward to confirm that $y, y' \in Y_2^w[t_2', f]$. Fix $\phi_2 \in \Delta(T_1 \times T_1)$. Then

$$\begin{aligned}
&\sum_{\tilde{t}_1, \tilde{t}_1'} \phi_2(\tilde{t}_1, \tilde{t}_1') u_2(y(\tilde{t}_1'), (\tilde{t}_1, t_2')) \\
= \;& \phi_2(t_1, t_1) u_2\big(y(t_1), (t_1, t_2')\big) + \phi_2(t_1, t_1') u_2\big(y(t_1'), (t_1, t_2')\big) + \phi_2(t_1, t_1'') u_2(y(t_1''), (t_1, t_2')) \\
& + \phi_2(t_1', t_1) u_2\big(y(t_1), (t_1', t_2')\big) + \phi_2(t_1', t_1') u_2\big(y(t_1'), (t_1', t_2')\big) + \phi_2(t_1', t_1'') u_2(y(t_1''), (t_1', t_2')) \\
& + \phi_2(t_1'', t_1) u_2\big(y(t_1), (t_1'', t_2')\big) + \phi_2(t_1'', t_1') u_2\big(y(t_1'), (t_1'', t_2')\big) + \phi_2(t_1'', t_1'') u_2(y(t_1''), (t_1'', t_2')) \\
= \;& \frac{3}{2}(\phi_1(t_1, t_1) + \phi_1(t_1, t_1') + \phi_1(t_1, t_1'')) + \frac{3}{2}(\phi_2(t_1', t_1) + \phi_2(t_1', t_1') + \phi_2(t_1', t_1'')) \\
& + 2(\phi_2(t_1'', t_1) + \phi_2(t_1'', t_1') + \phi_2(t_1'', t_1''))
\end{aligned}$$

whereas

$$\begin{aligned}
&\sum_{\tilde{t}_1, \tilde{t}_1'} \phi_2(\tilde{t}_1, \tilde{t}_1') u_2(y'(\tilde{t}_1'), (\tilde{t}_1, t_2')) \\
= \;& \phi_2(t_1, t_1) u_2\big(y'(t_1), (t_1, t_2')\big) + \phi_2(t_1, t_1') u_2\big(y'(t_1'), (t_1, t_2')\big) + \phi_2(t_1, t_1'') u_2(y'(t_1''), (t_1, t_2')) \\
& + \phi_2(t_1', t_1) u_2\big(y'(t_1), (t_1', t_2')\big) + \phi_2(t_1', t_1') u_2\big(y'(t_1'), (t_1', t_2')\big) + \phi_2(t_1', t_1'') u_2(y'(t_1''), (t_1', t_2')) \\
& + \phi_2(t_1'', t_1) u_2\big(y'(t_1), (t_1'', t_2')\big) + \phi_2(t_1'', t_1') u_2\big(y'(t_1'), (t_1'', t_2')\big) + \phi_2(t_1'', t_1'') u_2(y'(t_1''), (t_1'', t_2')) \\
= \;& \phi_2(t_1, t_1) + \phi_2(t_1, t_1') + \phi_2(t_1, t_1'').
\end{aligned}$$

This implies that for any $\phi_2 \in \Delta(T_1 \times T_1)$,

$$\sum_{\tilde{t}_1, \tilde{t}_1'} \phi_2(\tilde{t}_1, \tilde{t}_1') u_2(y(\tilde{t}_1'), (\tilde{t}_1, t_2')) > \sum_{\tilde{t}_1, \tilde{t}_1'} \phi_2(\tilde{t}_1, \tilde{t}_1') u_2(y'(\tilde{t}_1'), (\tilde{t}_1, t_2')).$$

Thus, the requirement for weak NWR is satisfied.

We therefore conclude that f satisfies weak NWR. □

We now show that the SCF f is implementable in interim rationalizable strategies.

Claim 7.6. *The SCF f is implementable in interim rationalizable strategies by the canonical mechanism we used in Theorem 6.3.*

Proof. We have shown that the SCF f satisfies weak IRM and weak NWR. Thus, by Theorem 6.3, f is implementable in interim rationalizable strategies by the canonical mechanism used in the proof of the theorem. □

We further claim that there are no mixed Bayesian equilibria in that canonical mechanism.

Claim 7.7. *There are no mixed Bayesian equilibria in the canonical mechanism implementing the SCF f used in Theorem 6.3.*

Proof. Since the SCF f fails BM, which is a necessary condition for Bayesian implementation, in particular it cannot be implemented in equilibrium (mixed or pure) by our canonical mechanism.[12] But every strategy profile induced by an equilibrium is rationalizable. Therefore, if there are any equilibria in the canonical mechanism, their outcome should be the SCF (because the SCF is implemented in rationalizable strategies). It then follows that the reason for the failure of implementation in Bayesian equilibrium by the canonical mechanism is that it does not have any equilibria in mixed or pure strategies. □

To illustrate the fact that there are no mixed Bayesian equilibria in the canonical mechanism, we consider the following strategy profile σ where $\sigma_i(t_i) = (m_i^1, m_i^2, m_i^3, m_i^4)$:

- $m_i^1[i] = t_i$ (i.e., each player announces her own type truthfully)
- $m_1^1[2] = t_2$ 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)
- $m_1^2 = m_2^2 = 1$ (i.e., each player announces 1 in the second component of the message)

By Step 1 of the proof of Theorem 6.3, every rationalizable strategy profile induces Rule 1. By construction, the strategy profile σ induces Rule 1. In Step 3 of the proof of Theorem 6.3, 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 $y : T_2 \to \Delta(A)$ such that $y(t_2) = y(t_2') = 0.99 \times a + 0.01 \times b$. Then, we obtain

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

[12] See Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989), and Jackson (1991) for the necessity of BM for implementation in pure Bayesian equilibrium, and Serrano and Vohra (2010) and Kunimoto (2019) for the necessity of mixed BM for implementation in mixed Bayesian equilibrium. Note that mixed BM is a strictly stronger condition than BM, as shown in Example 1 of Serrano and Vohra (2010).

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

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

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

In this section, we briefly discuss and raise other questions that are connected to our work.

8.1 Finite Mechanisms

Bergemann and Morris (2008) shows that, if an SCF f is implementable in rationalizable strategies by a finite mechanism, it satisfies IRM. Therefore, it follows as a simple corollary of our Lemma 5.8 that, if an SCF f is implementable in rationalizable strategies by a finite mechanism, it satisfies BM. It also follows that, if an SCF satisfies weak IRM but not IRM (as in Example 7.1), the SCF could be implemented in rationalizable strategies, but the implementing mechanism could never be finite.

For complete information environments, Chen *et al.* (2020b) characterizes rationalizable implementation by means of finite mechanisms when lotteries and transfers are allowed. The characterization is in terms of Maskin monotonicity*, a strenghthening of Maskin monotonicity.[13] Chen *et al.* (2020a) also shows, in the same environments, that Maskin monotonicity is necessary and sufficient for Nash implementation in 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 (2020), using finite mechanisms, robust implementation in rationalizable strategies and in interim equilibria are equivalent.

[13]This condition features in Bergemann *et al.* (2011) for the rationalizable implementation of SCFs, albeit allowing general mechanisms.

8.2 Complete information environments

Example 7.1 shows that rationalizable implementation could be more permissive than equilibrium implementation. Interestingly, this relation reverses in complete information environments for SCFs, that is, equilibrium implementation of SCFs is more permissive than rationalizable implementation in complete information environments. Bergemann *et al.* (2011) show that the necessary condition for rationalizable implementation is stronger than Maskin monotonicity, which is necessary for Nash implementation (Maskin, 1999). They also give an example of an SCF that is implementable in Nash equilibrium but not in rationalizable strategies. Xiong (2018) provides a complete characterization of SCFs that are implementable in rationalizable strategies. For the sufficiency part of the argument, he 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. Indeed, as shown in Kunimoto and Serrano (2019), when it comes to multi-valued social choice correspondences, rationalizable implementation is more permissive than equilibrium implementation in complete information environments.

9 Concluding Remarks

We have proposed weak interim rationalizable monotonicity (IRM) as a novel condition and showed that it is a necessary and almost sufficient condition for interim rationalizable implementation of social choice functions. We also show by means of an example 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. We plan to generalize the findings in this paper to multi-valued social choice rules, i.e., social choice sets, in a separate paper. We conclude the paper with mentioning two open questions left for future research.

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,

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

where $\Sigma = \Sigma_1 \times \cdots \times \Sigma_n$ and $\Sigma_i = \{\sigma_i |\ \sigma_i : T_i \to \Delta(M_i)\}$. 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 9.1. A mechanism Γ *doubly implements* an SCF f in Bayesian equilibria and rationalizable strategies if there exists an SCF $\tilde{f} \approx f$ 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 $g(m) = \hat{f}(t)$.

As we argue in the example of Section 7 that IRM is not necessary for interim rationalizable implementation and our canonical mechanism exploits the feature that there are no mixed Bayesian equilibria, one could investigate the connections between IRM and double implementation.

Responsive SCFs: An SCF f is responsive if, for all $i \in I$ and $t_i, t_i^{'} \in T_i$: $t_i \neq t_i^{'} \Rightarrow t_i \not\sim_i^f t_i^{'}$. Otherwise, f is nonresponsive. Then, for a responsive SCF, one could also investigate whether weak IRM and IRM are identical conditions. It is possible that the global inequalities embodied in the definition of responsiveness leave room to translate weak refutability into strong refutability, which makes weak IRM and IRM equivalent.

Appendix

Proof of Lemma 5.5: 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 $t_i^{'} \not\sim_i^f t_i$ 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 $y \in \bigcap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such that

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

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 $y^{'} \in \bigcap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$ such

that

$$\sum_{t_{-i},\tilde{t}_{-i}} \phi_i^{'}(t_{-i},\tilde{t}_{-i}) u_i\big(y^{'}(\tilde{t}_{-i}),(t_i,t_{-i})\big) > \sum_{t_{-i},\tilde{t}_{-i}} \phi_i^{'}(t_{-i},\tilde{t}_{-i}) u_i\big(f(t_i^{'},\tilde{t}_{-i}),(t_i,t_{-i})\big). \tag{7}$$

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 unresponsive 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 (7) that

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

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 $t_i^{'} \not\sim_i^f t_i$ 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 unresponsive 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

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

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 unresponsive 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

$$\sum_{t_{-i},\tilde{t}} \psi_i^{'}(t_{-i},\tilde{t}) u_i\big(f^{''}(\tilde{t}),(t_i,t_{-i})\big) > \sum_{t_{-i},\tilde{t}} \psi_i^{'}(t_{-i},\tilde{t}) u_i\big(f(t_i^{'},\tilde{t}_{-i}),(t_i,t_{-i})\big). \tag{8}$$

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 unresponsive 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, $y \in \bigcap_{\tilde{t}_i \in T_i} Y_i[\tilde{t}_i, f]$. Moreover, it follows from (8) that

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

Thus, f satisfies IRM. □

Proof of Lemma 6.2: 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

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

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

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

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 $y^{'\epsilon}$. By construction, y^{ϵ} and $y^{'\epsilon}$ 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 1, we have

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

We fix any such sufficiently large ϵ.

Third, since $\Delta^*(A)$ 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

$$\sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i})u_i\big(y^{z}(t'_{-i}),(t_i,t_{-i})\big) > \sum_{t_{-i},t'_{-i}} \phi_i(t_{-i},t'_{-i})u_i\big(y'^{z}(t'_{-i}),(t_i,t_{-i})\big). \tag{9}$$

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

Lastly, since $y_i^{t_i,f}$, 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

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

then it must be that

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

which contradicts (9).

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 $\bar{\alpha}$ assigns a positive weight to all $a \in A$, if

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

then it must be that

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

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

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

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

References

[1] Abreu, D. and H. Matsushima (1992), "Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information," Mimeo, Princeton University.

[2] Artemov, G., T. Kunimoto, and R. Serrano (2013), "Robust Virtual Implementation: Toward a Reinterpretation of the Wilson Doctrine," *Journal of Economic Theory* vol. 148, 424-447.

[3] Battigalli, P. and M. Siniscalchi (2003), "Rationalization and Incomplete Information," *The BE Journal of Theoretical Economics* vol. 3 (Advances), Article 3.

[4] Battigalli, P., A. Di Tillio, E. Grillo, and A. Penta (2011), "Interactive Epistemology and Solution Concepts in Games with Asymmetric Information," *The BE Journal of Theoretical Economics* vol. 11(Advances), Article 6.

[5] Bergemann, D. and S. Morris (2008), "Interim Rationalizable Implementation," Mimeo.

[6] Bergemann, D. and S. Morris (2009), "Robust Virtual Implementation," *Theoretical Economics*, vol. 4, 45-88.

[7] Bergemann, D., S. Morris, and O. Tercieux (2011), "Rationalizable Implementation," *Journal of Economic Theory*, vol. 146, 1253-1274.

[8] Bernheim, D. (1984), "Rationalizable Strategic Behavior." *Econometrica* vol. 52, 1007-1028.

[9] Brandenburger, A. and E. Dekel (1987), "Rationalizability and Correlated Equilibria," *Econometrica* vol. 55, 1391-1402.

[10] Chen, Y-C, T. Kunimoto, Y. Sun, and S. Xiong (2020a), "Maskin Meets Abreu and Matsushima," Mimeo, Singapore Management University.

[11] Chen, Y-C, T. Kunimoto, Y. Sun, and S. Xiong (2020b), "Rationalizable Implementation in Finite Mechanisms," Mimeo, Singapore Management University.

[12] Dekel, E., D. Fudenberg, and S. Morris (2007), "Interim Correlated Rationalizability," *Theoretical Economics*, vol. 2, 15-40.

[13] Di Tillio, A. (2011), "A Robustness Result for Rationalizable Implementation," *Games and Economic Behavior*, vol. 72, 301-305.

[14] Jackson, M. (1991), "Bayesian Implementation," *Econometrica*, vol. 59, 461-477.

[15] Jain, R. (2020), "Rationalizable Implementation of Social Choice Correspondences," Working Paper, Academia Sinica.

[16] Kunimoto, T. (2019), "Mixed Bayesian Implementation in General Environments," *Journal of Mathematical Economics*, vol. 82, 247-263.

[17] Kunimoto, T. and R. Saran (2020), "Robust Implementation in Rationalizable Strategies in General Mechanisms," Working Paper, Singapore Management University.

[18] Kunimoto, T. and R. Serrano (2019), "Rationalizable Implementation of Correspondences," *Mathematics of Operations Research*, vol. 44, 1326-1344.

[19] Lipman, B. (1994), "A Note on the Implications of Common Knowledge of Rationality," *Games and Economic Behavior*, vol. 6, 114-129.

[20] Maskin, E. (1999), "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies*, vol. 66, 23-38.

[21] Mezzetti, C. and L. Renou (2012), "Implementation in Mixed Nash Equilibrium," *Journal of Economic Theory*, vol. 147, 2357-2375.

[22] Oury, M. and O. Tercieux (2012), "Continuous Implementation," *Econometrica*, vol. 80, 1605-1637.

[23] Palfrey, T. and S. Srivastava (1989), "Implementation with Incomplete Information in Exchange Economies," *Econometrica*, vol. 57, 115-134.

[24] Pearce, D. (1984), "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* vol. 52, 1029-1050.

[25] Postlewaite, A. and D. Schmeidler (1986), "Implementation in Differential Information Economies," *Journal of Economic Theory*, vol. 39, 14-33.

[26] Serrano, R. and R. Vohra (2001), "Some Limitations of Virtual Bayesian Implementation," *Econometrica*, vol. 69, 785-792.

[27] Serrano, R. and R. Vohra (2005), "A Characterization of Virtual Bayesian Implementation," *Games and Economic Behavior*, vol. 50, 312-331.

[28] Serrano, R. and R. Vohra (2010), "Multiplicity of Mixed Equilibria in Mechanisms: A Unified Approach to Exact and Approximate Implementation," *Journal of Mathematical Economics*, vol. 46, 775-785.

[29] Xiong, S. (2018), "Rationalizable Implementation I: Social Choice Functions," Working Paper.