Econometrica, Vol. 90, No. 5 (September, 2022), 2437–2469

STRUCTURAL RATIONALITY IN DYNAMIC GAMES

MARCIANO SINISCALCHI
Economics Department, Northwestern University

The analysis of dynamic games hinges on assumptions about players' actions and beliefs at information sets that are not expected to be reached during game play. Under the standard notion of sequential rationality, these assumptions cannot be tested on the basis of observed, on-path behavior. This paper introduces a novel optimality criterion, *structural rationality*, which addresses this concern. In any dynamic game, structural rationality implies weak sequential rationality (Reny (1992)). If players are structurally rational, assumptions about on-path and off-path beliefs concerning off-path actions can be tested via suitable "side bets." Structural rationality also provides a theoretical rationale for the use of a novel version of the strategy method (Selten (1967)) in experiments.

KEYWORDS: Conditional probability systems, sequential rationality, strategy method.

1. INTRODUCTION

THE ANALYSIS OF DYNAMIC GAMES HINGES ON ASSUMPTIONS about players' actions and beliefs at information sets that are not expected to be reached during game play. A key aspect of Savage's (1954) foundational analysis of expected utility is to argue that the psychological notion of "belief" can and should be related to observable behavior. This paper introduces a notion of rationality in dynamic games that is just strong enough to permit the elicitation of beliefs, both on and off the predicted path of play. Moreover, in doing so, this paper introduces novel belief-elicitation techniques that broaden the range of predictions that can be tested experimentally.

In a single-person choice problem, the agent's beliefs can be elicited via "side bets" on the relevant events, with the stipulation that both the choice in the original problem and the side bets contribute to the overall payoff. Similarly, in a game with simultaneous moves, a player's beliefs can be elicited via side bets on opponents' actions (Luce and Raiffa (1957, Section 13.6)). [1]

However, in a dynamic game, the fact that certain information sets may be off the predicted path of play poses additional challenges. For instance, in the game of Figure 1 (cf. Van Damme (1989)), consider the subgame-perfect equilibrium profile (Out, (S, S)). Suppose first that an experimenter wishes to verify that, if Ann played In, Bob would indeed expect her to continue with S. If the simultaneous-move subgame was reached, the experimenter could offer Bob side bets on Ann's actions B versus S. However, Ann plays Out at the initial node in this equilibrium, so the subgame is never actually reached. Alternatively, the experimenter could try to elicit the *prior* probability that Bob assigns to Ann choosing In followed by S, and then update it by conditioning on the event that Ann

Marciano Siniscalchi: marciano@northwestern.edu
I thank three anonymous referees for their comments and suggestions. I also thank Amanda Friedenberg, as well as Pierpaolo Battigalli, Gabriel Carroll, Francesco Fabbri, Drew Fudenberg, Ben Golub, Julien Manili, Alessandro Pavan, Phil Reny, and participants at seminar presentations for helpful comments.

[1]For game-theoretic experiments implementing side bets see, for example, Nyarko and Schotter (2002), Costa-Gomes and Weizsäcker (2008), Rey-Biel (2009), and Blanco, Engelmann, Koch, and Normann (2010). For related approaches, see Aumann and Dreze (2009) and Gilboa and Schmeidler (2003).

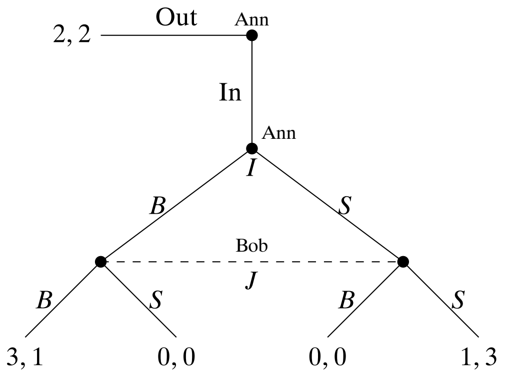

FIGURE 1.—The battle of the sexes with an outside option.

chooses In. However, in the equilibrium under consideration In has zero prior probability, so updating is not possible.

Now suppose that the experimenter wishes to verify that, at the beginning of the game, Ann believes that Bob would play S in the subgame. It would appear that offering Ann a side bet at the beginning of the game might work. However, in the equilibrium under consideration, Ann plays Out at the initial node; provided the side bet does not change her incentives (as it should not), Ann's own move prevents the subgame from being reached. Therefore, Ann understands that no side bet on Bob's move can actually be decided, or paid out. Thus, such a bet provides no real incentives to Ann. Again, elicitation fails—though for a different reason.

To address these issues, I propose the notion of *structural rationality*, which builds upon trembling-hand perfection (Selten (1975)). Fix a player's beliefs at each information set. A perturbation of the player's beliefs is a sequence of probabilities that assigns positive weight to each information set where the player moves, and approximates the player's conditional beliefs there. A strategy is structurally rational given the player's beliefs if it maximizes her ex ante expected payoff with respect to some perturbation. Thus, as in trembling-hand perfection, each player sees every information set as possible, if arbitrarily unlikely. However, unlike in trembling-hand perfection, different perturbations of the player's beliefs can justify different structural best replies. In this sense, structural rationality takes the *possibility* of surprises seriously, without committing to any specific "theory" about them.

Proposition 1 draws a connection between structural rationality and a notion of "robust" preference reminiscent of Bewley (2002). Theorem 1 shows that structural rationality implies weak sequential rationality (Reny (1992), Battigalli (1997), Battigalli and Siniscalchi (2002)).[2] The main result of this paper, Theorem 2, shows that, under structural rationality, side bets offered at the beginning of the game allow the incentive-compatible elicitation of beliefs at every information set, whether on or off the expected path of play.

This result leverages a (to the best of my knowledge) novel experimental design in which all players are asked to *report* their intended strategy, and are rewarded if their actual play conforms to their report. This is a variant of the *strategy method* of Selten

[2]Theorem 3 in Appendix B.3 provides a partial converse: under suitable "genericity" assumptions on payoffs at terminal histories, if a strategy is weakly sequentially for given beliefs, that strategy is also structurally rational.

(1967), which requires that players commit to (rather than just announce) extensive-form strategies. Structural rationality ensures that players have strict incentives to report the strategy that they are in fact planning to follow. Side bets are then paid out on the basis of reported strategies, which are always observed. To illustrate, in the game in Figure 1, this design gives Bob strict incentives to bet on Ann playing S in the subgame: even if he expects Ann to play Out, by structural rationality Bob will take seriously the possibility of being surprised, and will bet accordingly. Similarly, Ann has strict incentives to bet on Bob playing S in the subgame: even if she herself plans to (and indeed will) play Out, she recognizes that—by structural rationality—Bob will plan on playing S, and will report this truthfully, so her own bet will be paid out accordingly. Structural rationality is crucial to these conclusions; see Example 4.

The companion paper Siniscalchi (2022) provides an axiomatic analysis of the notion of "robust preference" that underlies structural rationality, and shows that it is the most permissive such notion that still allows the identification of beliefs and utilities. A second paper, Siniscalchi (in preparation), shows how to incorporate it into different solution concepts. Section 6.5 in the present paper takes a first step and defines a version of sequential equilibrium in which structural rationality is the notion of best reply. It also draws a connection with trembling-hand perfection.

Organization. Section 2 introduces the required notation. Section 3 formalizes beliefs and sequential rationality. Section 4 defines structural rationality. Section 5 contains the main results. Section 6 provides additional discussion and extensions. All proofs are in the Appendix.

2. BASIC NOTATION

Following Osborne and Rubinstein (1994, Definition 200.1, pp. 200–201), a finite dynamic game with imperfect information is represented by a tuple $(N, A, Z, P, (\mathcal{I}_i, u_i)_{i\in N})$, where:

- N is the set of *players* and A is the set of *actions*.
- $Z \subset \bigcup_{0\le t<\infty} A^t$ is the finite set of *terminal histories*. Given Z, $H \equiv \bigcup_{(a^1,\dots,a^t)\in Z}\{(a^1, \dots, a^\tau) : 0 \le \tau \le t\}$ is the set of all *histories*, including the *root* (empty history) ϕ.
- $P : H \setminus Z \to N$ is the *player function*.
- $\mathcal{I}_i$ is the collection of *information sets* of player i; it is a partition of $P^{-1}(\{i\})$, and is such that, if $(a_1, \dots, a_K), (b_1, \dots, b_L) \in I$ for some $I \in \mathcal{I}_i$, and $(a_1, \dots, a_K, a) \in H$, then $(b_1, \dots, b_L, a) \in H$. That is, the same actions are available at every history in the same information set.
- $u_i : Z \to \mathbb{R}$ is the *payoff function* for player i.

Section 6.1 shows how to allow for incomplete information.

The analysis in this paper mostly focuses on the following derived objects:

- For every $i \in N$ and $I \in \mathcal{I}_i$, $A(I) = \{a \in A : \exists(a_1, \dots, a_k) \in I, (a_1, \dots, a_k, a) \in H\}$ is the (nonempty) set of *actions available to i at I*.[3]
- For every $i \in N$, $S_i = \prod_{I\in\mathcal{I}_i} A(I)$ is the set of *strategies* of player i; the action specified by $s_i \in S_i$ at $I \in \mathcal{I}_i$ is denoted $s_i(I)$, and as usual $S = \prod_{i\in N} S_i$ and $S_{-i} = \prod_{j\neq i} S_j$.
- For every $h = (a_1, \dots, a_K) \in H$, $S(h) = \{s \in S : \forall k = 1, \dots, K, \exists i \in N, I \in \mathcal{I}_i \text{ s.t. } (a_1, \dots, a_{k-1}) \in I, a^k = s_i(I)\}$ is the set of strategy profiles that *induce* h. Let $S_i(h) = \text{proj}_{S_i} S(h)$ and $S_{-i}(h) = \text{proj}_{S_{-i}} S(h)$.

[3] This is well posed by the assumption that the same actions are available at every $h \in I$.

- For every $i \in N$ and $I \in \mathcal{I}_i$, $S(I) = \bigcup_{h\in I} S(h)$ is the set of strategy profiles that *induce I*. Let $S_i(I) = \text{proj}_{S_i} S(I)$ and $S_{-i}(I) = \text{proj}_{S_{-i}} S(I)$. If $s_{-i} \in S_{-i}(I)$, say that s_{-i} *allows I*.[4]
- The *strategic-form payoff function* of player $i \in N$ is $U_i : S_i \times S_{-i} \to \mathbb{R}$, defined by $U_i(s_i, s_{-i}) = u_i(z)$ for all $z \in Z$ and $(s_i, s_{-i}) \in S(z)$.

As usual, for any $s_i \in S_i$ and $p \in \Delta(S_{-i})$, let $U_i(s_i, p) = \sum_{s_{-i}} U_i(s_i, s_{-i}) \cdot p(\{s_{-i}\})$; and for any $\sigma_i \in \Delta(S_i)$, let $U_i(\sigma_i, p) = \sum_{t_i\in S_i} \sigma_i(t_i)U_i(t_i, p)$.

Sets of the form $S_{-i}(I)$, for $I \in \mathcal{I}_i$, are called *conditioning events*. In preparation for Definition 1, it is convenient to define $S_{-i}(\phi) = S_{-i}$ for *all* players $i \in N$, not just $i = P(\phi)$.

I assume that the game has *perfect recall*, analogously to Definition 203.3 in Osborne and Rubinstein (1994); see Appendix A. This has two implications that are used in the analysis. First, for every $i \in N$ and $I \in \mathcal{I}_i$, $S(I) = S_i(I) \times S_{-i}(I)$. Second, the set $S(I)$ satisfies *strategic independence* (Mailath, Samuelson, and Swinkels (1993, Definition 2 and Theorem 1)): for every $s_i, t_i \in S_i(I)$, there is $r_i \in S_i(I)$ such that $U_i(r_i, s_{-i}) = U_i(t_i, s_{-i})$ for all $s_{-i} \in S_{-i}(I)$, and $U_i(r_i, s_{-i}) = U_i(s_i, s_{-i})$ for all $s_{-i} \in S_{-i} \setminus S_{-i}(I)$. Intuitively, r_i is the strategy that coincides with s_i everywhere except at I and all subsequent information sets, where it coincides with t_i.

3. BELIEFS AND WEAK SEQUENTIAL RATIONALITY

Throughout the remainder of this paper, unless referring to a specific example, I fix an arbitrary dynamic game $(N, A, Z, P, (\mathcal{I}_i, u_i)_{i\in N})$.

I represent player i's beliefs as a collection of probability distributions over coplayers' strategies, indexed by her information sets $I \in \mathcal{I}_i$;[5] compare Rényi (1955), Myerson (1986), Ben-Porath (1997), Kohlberg and Reny (1997), Battigalli and Siniscalchi (2002). It is also convenient to assume that every player has a prior belief, even if she does not move at the root ϕ of the game. The probabilities $(\mu(\cdot|I))_{I\in\mathcal{I}_i\cup\{\phi\}}$ have a dual interpretation. From an *interim* perspective, every $\mu(\cdot|I)$ can be interpreted as the beliefs that player i would hold upon reaching I. This is the interpretation that best fits the notion of sequential rationality. Alternatively, the entire probability array $(\mu(\cdot|I))_{I\in\mathcal{I}_i\cup\{\phi\}}$ can be viewed as a description of player i's *prior* beliefs, according to which every information set is reached with positive, but possibly "infinitesimal" probability. In this interpretation, $\mu(\{s_{-i}\}|I)$ describes the likelihood of strategy profile s_{-i} relative to that of information set I, which may itself be infinitely unlikely a priori. This interpretation is particularly apt from the perspective of structural rationality.

DEFINITION 1: An array $\mu = (\mu(\cdot|I))_{I\in\mathcal{I}_i\cup\{\phi\}} \in \Delta(S_{-i})^{\mathcal{I}_i\cup\{\phi\}}$ is a *consistent conditional probability system (CCPS)* for player i if there is a sequence $(p^k)_{k\geq 1} \in \Delta(S_{-i})^{\mathbb{N}}$ such that, for all $I \in \mathcal{I}_i \cup \{\phi\}$, $p^k(S_{-i}(I)) > 0$ for all $k \geq 1$, and $\lim_{k\to\infty} p^k(\cdot|S_{-i}(I)) = \mu(\cdot|I)$. Such a sequence (p^k) is called a *perturbation* of μ. Denote the set of CCPSs for player i by $\Delta(S_{-i}, \mathcal{I}_i)$.

Adapting arguments in Myerson (1986), one readily sees that a CCPS is a "conditional probability system" à la Rényi (1955). However, it satisfies further restrictions; see Siniscalchi (2022).

[4]That is, if i's coplayers follow the profile s_{-i}, I *can* be reached; whether it *is* reached depends upon i's play.
[5]Definition 1 implies that, equivalently, one can take the corresponding conditioning events $S_{-i}(I)$ as indices.

The probabilities p^k in Definition 1 need *not* have full support. In particular, in games with simultaneous moves, the constant sequence defined by $p^k = \mu(\cdot|\phi)$ for all k is a perturbation of a player's (trivial) CCPS $\mu = \mu(\cdot|\phi)$.

To formalize sequential rationality, I follow Reny (1992) and Rubinstein (1991), and only require that a strategy s_i of player i be optimal at information sets that s_i allows. Optimality at other information sets is best viewed as a description of other players' (equilibrium) beliefs about i, rather than part of player i's decision-making. Following Reny (1992), I call this notion "weak sequential rationality," to distinguish it from the definition in Kreps and Wilson (1982).

DEFINITION 2: Fix a CCPS $\mu \in \Delta(S_{-i}, \mathcal{I}_i)$. A strategy $s_i \in S_i$ is *weakly sequentially rational given* μ if, for every $I \in \mathcal{I}_i \cup \{\phi\}$ with $s_i \in S_i(I)$, and all $t_i \in S_i(I)$, $U_i(s_i, \mu(\cdot|I)) \geq U_i(t_i, \mu(\cdot|I))$.

4. STRUCTURAL RATIONALITY

For conciseness, all definitions and results in this section apply to a player $i \in N$, and a CCPS $\mu \in \Delta(S_{-i}, \mathcal{I}_i)$ for player i in the dynamic game $(N, A, Z, P, (\mathcal{I}_i, u_i)_{i\in N})$.

DEFINITION 3: A strategy $s_i \in S_i$ is *structurally rational given* μ if there is a perturbation $(p^k)_{k\geq 1}$ of μ such that, for every $t_i \in S_i$, $U_i(s_i, p^k) \geq U_i(t_i, p^k)$ for all $k \geq 1$.

Structural rationality depends upon (i) the extensive-form structure of the game, and specifically on the collection $\{S_{-i}(I) : I \in \mathcal{I}_i \cup \{\phi\}\}$ of conditioning events; and (ii) on player i's entire CCPS. Conditioning events and the associated conditional beliefs characterize the set of perturbations. Hence, structural rationality is not invariant with respect to the strategic form.

That said, in simultaneous-move ("strategic-form") games, one particular perturbation of μ is given by $p^k = \mu(\cdot|\phi)$ for all k. This is also the case in general dynamic games, if player i's prior $\mu(\cdot|\phi)$ assigns positive probability to every $I \in \mathcal{I}_i$. By Definition 3, *in these cases, a strategy is structurally rational given* μ *if and only if maximizes player i's ex ante expected payoff*.

EXAMPLE 1: In the game of Figure 1, suppose Bob's CCPS μ reflects his beliefs in the subgame-perfect equilibrium (Out, (S, S)), so $\mu(\{\text{Out}\}|\phi) = 1$ and $\mu(\{\text{In}S\}|J) = 1$. Any perturbation $(p^k)_{k\geq 1}$ of μ must satisfy $p^k(S_a(J)) = p^k(\{\text{In}S, \text{In}B\}) > 0$ for each k, and $p^k(\{\text{In}S\}|S_a(J)) \to 1$. For k large enough, $U_b(S, p^k) > U_b(B, p^k)$, so S is the only structurally rational strategy given μ.

EXAMPLE 2: In Figure 2, Ann's beliefs μ satisfy $\mu(\{d\}|\phi) = \mu(\{a\}|I) = 1$, so any perturbation $(p^k)_{k\geq 1}$ of μ must satisfy $p^k(S_{-a}(I)) = p^k(\{a\}) > 0$ for each k and $p^k(\{d\}) = p^k(\{d\}|S_b(\phi)) \to 1$.

Denote by D_1 either one of the realization-equivalent strategies D_1D_2, D_1A_2. If $x < 2$, eventually $U_a(D_1, p^k) > U_a(s_a, p^k)$ for any strategy $s_a \neq D_1$ of Ann, so D_1 is the unique structurally rational strategy given μ. If instead $x = 2$, $U_a(A_1D_2, p^k) > U_a(s_a, p^k)$ for all k and all $s_a \neq A_1D_1$. Thus, A_1D_2 is the unique structurally rational strategy given μ. By comparison, for $x = 2$, both D_1 and A_1D_2 are weakly sequentially rational given μ.

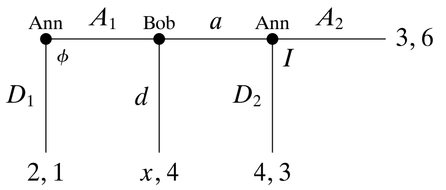

FIGURE 2.—A centipede-like game parameterized by $x \in [0, 2]$.

EXAMPLE 3: The game in Figure 3 is an extension of "Matching Pennies." Denote Ann's CCPS by μ, and assume that, as in the unique subgame-perfect equilibrium of this game, Ann initially expects Bob to play h and t with probability $\frac{1}{2}$: $\mu(\{h\}|\phi) = \mu(\{t\}|\phi) = \frac{1}{2}$. Denote by T any one of the realization-equivalent strategies of Ann that choose T at ϕ.

Any perturbation $(p^k)_{k\geq 1}$ of μ must satisfy $p^k(\{o\}) > 0$, $p^k(\{h\}) \to \frac{1}{2}$, and $p^k(\{t\}) \to \frac{1}{2}$. Since $p^k(\{o\}) > 0$ implies $U_a(\text{HL}, p^k) > U_a(\text{HR}, p^k)$, HR is not structurally rational given μ. If $2p^k(\{o\}) + p^k(\{h\}) > -p^k(\{o\}) + p^k(\{t\})$, then $U_a(\text{HL}, p^k) > U_a(T, p^k)$, If however $2p^k(\{o\}) + p^k(\{h\}) < -p^k(\{o\}) + p^k(\{t\})$, then $U_a(\text{HL}, p^k) < U_a(T, p^k)$. Thus, both HL and T are structurally rational given μ. This illustrates the *robustness* aspect of Definition 3: since all perturbations of Ann's beliefs μ are allowed, both HL and T are structurally rational given μ.

5. MAIN RESULTS

5.1. *Bewley-Style Characterization*

Structural rationality admits a characterization via a notion of "robust preference" in the spirit of Bewley's (2002) theory of Knightian uncertainty.

PROPOSITION 1: *A strategy $s_i \in S_i$ is structurally rational given μ if and only if there is no $\sigma_i \in \Delta(S_i)$ such that, for every perturbation $(p^k)_{k\geq 1}$ of μ, eventually $U_i(\sigma_i, p^k) > U_i(s_i, p^k)$.*

Thus, if s_i is *not* structurally rational, there is a mixed strategy σ_i that is "robustly better" than s_i; that is, σ_i eventually yields strictly higher expected payoff than s_i against *all* perturbations of μ. The companion paper Siniscalchi (2022) axiomatizes this robust preference relation.

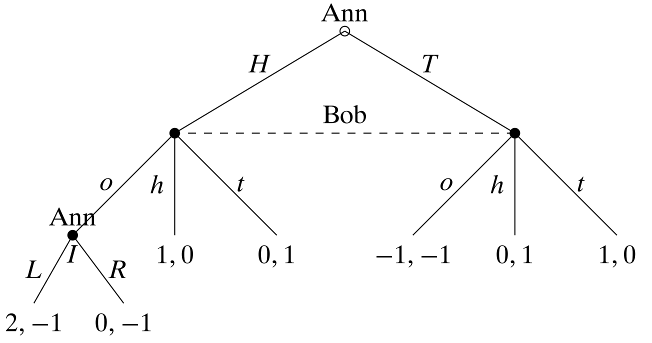

FIGURE 3.—Modified matching pennies.

5.2. *Structural and Weak Sequential Rationality*

THEOREM 1: *Fix a player $i \in N$ and a CCPS $\mu \in \Delta(S_{-i}, \mathcal{I}_i)$ for i. If strategy $s_i \in S_i$ is structurally rational given μ, then it is weakly sequentially rational given μ.*

The proof of Theorem 1 shows that, if s_i is structurally rational, it must specify an optimal continuation against some *perturbed* conditional belief $p^k(\cdot|S_{-i}(I)) \to \mu(\cdot|I)$ at every information set I with $s_i \in S_i(I)$. By way of contrast, weak sequential rationality only requires optimality against the limiting beliefs $\mu(\cdot|I)$. This is reminiscent of the difference between extensive-form trembling-hand perfect and sequential equilibrium (Kreps and Wilson (1982)), or between strategic-form perfect equilibrium and weak sequential equilibrium (Reny (1992)). Leveraging "generic equivalence" results from the cited papers, one can show that, for generic assignments of payoffs at terminal histories, in almost every (weakly) sequential equilibrium, every strategy played with positive probability is structurally rational.

This conclusion is not quite a "generic converse" to Theorem 1. The key limitation is that (weak) sequential equilibrium employs a more restrictive notion of beliefs than those allowed in Definition 1 and Theorem 1, namely "consistent assessments" à la Kreps and Wilson (1982) (see also Section 6.5). Theorem 3 in Appendix B.3 provides a proper "generic converse," using a notion of genericity that can be verified directly, by inspecting payoffs at terminal histories.

5.3. *Eliciting Conditional Beliefs*

Finally, I leverage structural rationality to elicit players' beliefs. A key requirement is that, in eliciting a player's beliefs, one must not alter the other players' strategic incentives. This distinguishes belief elicitation in games from elicitation in decision problems.

I restrict attention to *binary bets*: each player i can either bet on the realization of an event $E_i \subseteq S_{-i}$ (e.g., "Ann plays InS" in Figure 1) conditional upon reaching a given information set $I_i \in \mathcal{I}_i$ (e.g., J), or receive a guaranteed payoff of $p_i \in [0, 1]$ "utils" if I_i is reached. As will be shown, player i's choice of bet (E_i or p_i) will reveal whether or not she assigns probability at least p_i to E_i given I_i. It is straightforward to adapt the approach introduced here to offer players a menu of bets, or alternative mechanisms (e.g., Becker, DeGroot, and Marschak (1964)).

The elicitation mechanism consists of two phases. In the first, each player i simultaneously chooses a *bet* (or "wager") $w_i \in \{E_i, p_i\}$ and an *reported strategy* $\bar{s}_i \in S_i$, and the experimenter randomly selects one of the players—henceforth, "the selected player." In the second phase, the selected player plays the original game with the experimenter, who faithfully implements the reported strategies of the other players.[6]

At each terminal history, players who were not selected receive a fixed payoff (say, 0 utils) independent of their choices in the first phase and of play in the second phase. The selected player i instead receives an equal-chance lottery over three prizes: a *direct-play* prize, equal to the payoff determined by the realized play in the second phase of the mechanism; a *betting* prize, which depends on her bet w_i and the *reported* strategies of the other players, $\bar{s}_{-i}$; and a *bonus* $\epsilon > 0$ if her direct play is consistent with her reported strategy $\bar{s}_i$.

[6]Alternatively, players may play *separately* with the experimenter, either simultaneously or sequentially, provided they do not observe each other's moves. However, the required notation is much more cumbersome.

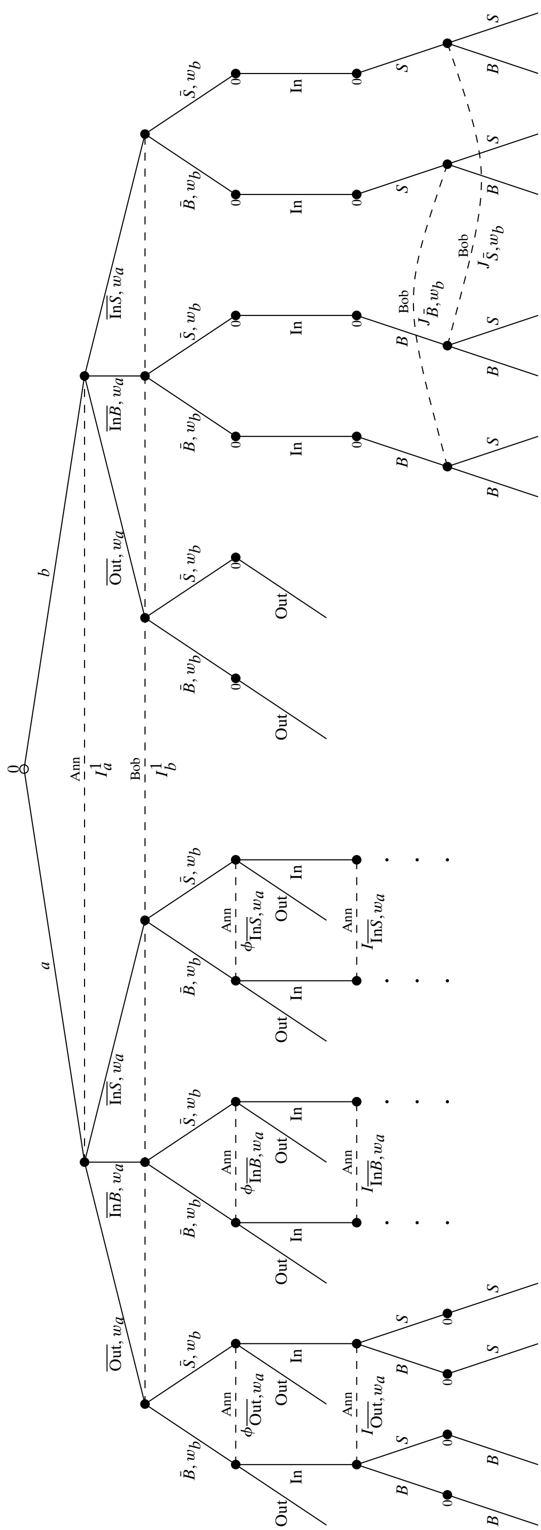

FIGURE 4.—A stylized representation of the elicitation game tree for Figure 1.

Figure 4 shows the game tree of the elicitation mechanism for the game in Figure 1, with one graphical simplification: each action in the first stage (e.g., $(\bar{B}, w_b)$ for Bob at information set I_b^1) actually represents *two* actions, one for each possible bet (e.g., $(\bar{B}, E_b)$ and $(\bar{B}, p_b)$).

I now formally define the elicitation game associated with an arbitrary dynamic game $(N, A, Z, P, (\mathcal{I}_i, u_i)_{i\in N})$. I allow for bets to be offered to any subset of players; this way the analysis will include a version of the strategy method (without elicitation) as a special case.

DEFINITION 4: A *questionnaire* is a collection $Q = (I_i, W_i)_{i\in N}$ such that, for every $i \in N$, $I_i \in \mathcal{I}_i$, and either $W_i = \{*\}$ or $W_i = \{E, p\}$ for some $E \subseteq S_{-i}(I_i)$ and $p \in [0, 1]$.[7]

Fixing a questionnaire Q, the sets of players and actions in the elicitation game are

$$N^* = N \cup \{0\} \quad \text{and} \quad A^* = N \cup \bigcup_{i\in N} (S_i \times W_i) \cup A. \tag{1}$$

Player 0 is the experimenter. Actions include the experimenter's choice of a selected player $i \in N$, and each subject i's choices of a reported strategy $\bar{s}_i \in S_i$ and bet $w_i \in W_i$.

Next, I define terminal histories $z^* \in Z^*$. In the first phase of the elicitation game, the experimenter moves first, then players move according to their index. In the second phase, the selected player n plays with the experimenter; the resulting sequence of actions must be a terminal history z in the original game. Along this history, whenever the player on the move is $j \neq n$, the experimenter faithfully carries out j's reported action. Formally, if the profile of reported strategies of players other than n is $\bar{s}_{-n}$, then $\bar{s}_{-n}$ must allow z. However, history z *need not* also be allowed by $\bar{s}_n$: regardless of her choice of reported strategy $\bar{s}_n$, the selected player can choose any course of action that is also available in the original game. Thus,

$$\begin{aligned} Z^* = \big\{ &\big(n, (\bar{s}_1, w_1), \dots, (\bar{s}_N, w_N), z\big) : \\ & n \in N, (\bar{s}_i, w_i) \in S_i \times W_i \; \forall i \in N, z \in Z, \bar{s}_{-n} \in S_{-i}(z) \big\}, \end{aligned} \tag{2}$$

where, consistently with Osborne and Rubinstein (1994), given two lists of actions $(a_1, \dots, a_L)$ and $(b_1, \dots, b_K) \equiv h$, I write $(a_1, \dots, a_L, h)$ to denote the joined list $(a_1, \dots, a_L, b_1, \dots, b_K)$.

As in Section 2, given the set Z^* of terminal histories, one can define the set H^* of all histories, terminal or not. With this, the player function is defined as

$$P^*(h^*) = \begin{cases} i & i \in N, h^* = \big(n, (\bar{s}_1, w_1), \dots, (\bar{s}_{i-1}, w_{i-1})\big), \\ P(h) & h^* = \big(n, (\bar{s}_1, w_1), \dots, (\bar{s}_N, w_N), h\big), h \notin Z, P(h) = n, \\ 0 & h^* = \phi^* \text{ or } h^* = \big(n, (\bar{s}_1, w_1), \dots, (\bar{s}_N, w_N), h\big), h \notin Z, P(h) \neq n. \end{cases} \tag{3}$$

Now turn to information. The experimenter has perfect information:

$$\mathcal{I}_0^* = \{\phi^*\} \cup \big\{\big\{\big(n, (\bar{s}_1, w_1), \dots, (\bar{s}_N, w_N), h\big)\big\} \subset H^* \setminus Z^* : P(h) \neq n\big\}. \tag{4}$$

[7]If $W_i = \{*\}$, I_i can be arbitrarily specified, and is immaterial.

In the first phase of the elicitation game, each player $i \in N$ does not observe the choices of those who moved before him: thus, her sole information set in the first phase is

$$I_i^1 = N \times \prod_{j=1}^{i-1} (S_j \times W_j) \subset H^*. \tag{5}$$

In the second phase, whenever the selected player i moves, she recalls her own reported strategy and bet, and *receives the same information as in the original game about other players' moves*—though these are carried out by the experimenter on their behalf. For instance, at $J_{\bar{B},w_b}$ in Figure 4, Bob observes In (and hence can infer that Ann's reported strategy is either $\overline{\text{In}B}$ or $\overline{\text{In}S}$). Thus, at $J_{\bar{B},w_b}$, Bob has the same information about Ann's prior move as at J in the game of Figure 1. To formalize this, for every $I \in \mathcal{I}_i$ and $(\bar{s}_i, w_i) \in S_i \times W_i$, let

$$I_{\bar{s}_i,w_i} = \{(n, (\bar{t}_1, v_1), \ldots, (\bar{t}_N, v_N), h) \in H^* : n = i, \bar{t}_i = s_i, \bar{v}_i = w_i, h \in I\}. \tag{6}$$

Then, for every player $i \in N$, the collection of information sets in the elicitation game is

$$\mathcal{I}_i^* = \{I_i^1\} \cup \{I_{\bar{s}_i,w_i} : I \in \mathcal{I}_i, (\bar{s}_i, w_i) \in S_i \times W_i\}. \tag{7}$$

Finally, payoffs are specified as follows: for all $z^* = (n, (\bar{s}_i, w_i)_{i\in N}, z) \in Z^*$,

$$u_i^*(z^*) = \begin{cases} 0 & i = 0 \text{ or } i \in N \setminus \{n\}, \\ \frac{1}{3}u_i(z) + \frac{1}{3}B(w_i, \bar{s}_{-i}) + \frac{1}{3} \cdot \epsilon \cdot 1_{\bar{s}_i \in S_i(z)} & i = n, \end{cases}$$

$$\text{where } B(E, \bar{s}_{-i}) = 1_{\bar{s}_{-i}\in E}, B(p, \bar{s}_{-i}) = p \cdot 1_{\bar{s}_{-i}\in S_{-i}(I_i)}, \text{ and } B(w_i, \bar{s}_{-i}) = 0 \text{ otherwise.} \tag{8}$$

For the selected player $i = n$, $u_i(z)$ is the *direct-play* payoff, $B(w_i, \bar{s}_{-i})$ is the *betting* payoff, and $\epsilon \cdot 1_{\bar{s}_i\in S_i(z)}$ is the *bonus*, paid out only if her direct play is consistent with her reported strategy.[8]

The complete definition of the elicitation game can now be stated.

DEFINITION 5: The *elicitation game for* $Q = (I_i, W_i)_{i\in N}$ *with bonus* ϵ is the tuple $(N^*, A^*, Z^*, P^*, (\mathcal{I}_i^*, u_i^*)_{i\in N\cup\{0\}}, \epsilon)$, where $\epsilon > 0$ and the other elements are as in equations (1)–(8).

How does the game thus defined allow the elicitation of beliefs—provided players are structurally rational? At a broad level, the mechanism works in three conceptual steps.

First, when selected to play directly, player n will choose a course of action that is part of a structurally rational strategy given her beliefs in the elicitation game. But, fixing n's choice of a reported strategy $\bar{s}_n$ and bet w_n, there is a one-to-one correspondence between information sets $I_{\bar{s}_n,w_n}$ in the second phase of the elicitation game and information sets I in the original game. Hence, if n's beliefs at $I_{\bar{s}_n,w_n}$ in the elicitation game "*agree* with" her beliefs at I in the original game, then any structurally rational course of action in

[8]In particular, in Figure 4, if $\bar{s}_b = \bar{B}$, Bob is selected, and Ann chooses $\bar{s}_a = \text{Out}$, the experimenter must play Out, so Bob's direct move is not observed. However, since intuitively there is "no evidence" that Bob would have deviated from her reported strategy, he still receives the bonus ϵ,

the former is structurally rational in the latter, and conversely. Thus, player n's strategic incentives are preserved.

Second, the selected player n's play in the second phase of the game is not limited by her choice of reported strategy $\bar{s}_n$. However, n *does* get a bonus if $\bar{s}_n$ is consistent with her direct play. Hence, at information set I_n^1, player n has an incentive to *correctly anticipate* her direct play, and report a strategy $\bar{s}_n$ that is consistent with it—not just on-path, but also following other players' unexpected actions. Moreover, by the previous argument, under belief agreement, her reported strategy $\bar{s}_n$ will also be consistent with her play in the *original* game.

Finally, suppose the experimenter wants to elicit the beliefs that another player i holds in the original game about n's moves. In the elicitation game, i bets on n's *reported* strategy. But, as was just argued, under belief agreement this is equivalent to betting on n's play in the original game. And since bets are always observed and paid out in the elicitation game, every player has (strict) incentives to bet in accordance with her beliefs.

To formalize this intuition, I first describe strategies in the elicitation game. Identify the set of strategies S_0^* for the experimenter with N, the set of players (at all other histories, the experimenter has a single available action). A strategy $s_i^* \in S_i^*$ for a player $i \in N$ must specify a reported strategy $\bar{s}_i$ and bet w_i at I_i^1. In addition, it must specify an action at *every* information set in the second phase of the elicitation game, including those that do *not* follow i's actual choice of $\bar{s}_i$ and w_i at I_i^1, and are thus not payoff-relevant. To focus on the payoff-relevant components of strategies, for each player $i \in N$, define the "reported-strategy" map $\mathbf{r}_i : S_i^* \to S_i$, bet or "wager" map $\mathbf{w}_i : S_i^* \to W_i$, and "direct-play" map $\mathbf{d}_i : S_i^* \to S_i$ as follows: for every $s_i^* \in S_i^*$, if $s_i^*(I_i^1) = (\bar{s}_i, w_i)$ then $\mathbf{r}_i(s_i^*) = \bar{s}_i$, $\mathbf{w}_i(s_i^*) = w_i$, and

$$\forall I \in \mathcal{I}_i, \quad \mathbf{d}_i\big(s_i^*\big)(I) = s_i^*(I_{\bar{s}_i, w_i}). \tag{9}$$

DEFINITION 6: Fix a player $i \in N$ and a CCPS $\mu^* \in \Delta(S_{-i}^*, \mathcal{I}_i^*)$. Say that μ^* *agrees with* $\mu \in \Delta(S_{-i}, \mathcal{I}_i)$ if, for every $s_{-i} \in S_{-i}$ and $n \in N$,

$$\mu^*\big(\big\{t_{-i}^* : t_0^* = n, \mathbf{r}_j\big(t_j^*\big) = s_j \ \forall j \in N \setminus \{i\}\big\} \big| \phi^*\big) = \frac{1}{N}\mu\big(\{s_{-i}\} \big| \phi\big), \tag{10}$$

$$\mu^*\big(\big\{t_{-i}^* : t_0^* = i, \mathbf{r}_j\big(t_j^*\big) = s_j \ \forall j \in N \setminus \{i\}\big\} \big| I_{\bar{s}_i, w_i}\big) = \mu\big(\{s_{-i}\} \big| I\big) \quad \forall I_{\bar{s}_i, w_i} \in \mathcal{I}_i^*. \tag{11}$$

Thus (i) ex ante, i believes that each player has an equal chance of being selected to play directly, and that the selection process is independent of coplayers' choices of reported strategies; and (ii) at every information set, i holds the same beliefs about each coplayer j's reported strategy as about j's strategy in the original game.[9]

More than one CCPS for player i in the elicitation game may agree with her CCPS in the original game, because i may assign different probabilities to her coplayers' choices of side bets. However, these differences do not affect i's payoff.

The main result of this section can now be stated: if players' beliefs about others' reported strategies are the same as in the original game, then (i) the elicitation mechanism does not change the set of structurally rational strategies, (ii) belief bounds can be elicited from initial, observable betting choices, and (iii) reported strategies are consistent with direct play.

[9]Parts (i) and (iii) in Theorem 2 suggest that one could alternatively define "agreement" as meaning that i believes that coplayers' direct play coincides with (a) their play in the original game, and (b) their reported strategies in the elicitation game. Doing so is possible, but notationally more cumbersome.

THEOREM 2: *Fix a questionnaire* $(I_i, W_i)_{i\in N}$ *and let* $(N^*, (S_i^*, \mathcal{I}_i^*, U_i^*)_{i\in N^*}, S^*(\cdot))$ *be the associated elicitation game. For any CCPS* $\mu_i \in \Delta(S_{-i}, \mathcal{I}_i)$ *for player* $i \in N$, *there is a CCPS* $\mu_i^* \in \Delta(S_{-i}^*, \mathcal{I}_i^*)$ *that agrees with* μ_i; *and for any such* μ_i^*, *and strategy* $s_i^* \in S_i^*$ *that is structurally rational for* μ_i^*:

(i) $\mathbf{r}_i(s_i^*)$ *and* $\mathbf{d}_i(s_i^*)$ *are structurally rational for* μ_i;

(ii) *if* $W_i = (E, p)$ *and and* $\mathbf{w}_i(s_i^*) = E$ (*resp*., $\mathbf{w}_i(s_i^*) = p$), *then* $\mu_i(E|I_i) \geq p$ (*resp*., $\mu_i(E|I_i) \leq p$);

(iii) *for all* $z \in Z$, $\mathbf{r}_i(s_i^*) \in S_i(z)$ *if and only if* $\mathbf{d}_i(s_i^*) \in S_i(z)$.

Conversely, for every $s_i \in S_i$ *that is structurally rational for* μ_i, *there is* $s_i^* \in S_i^*$ *with* $\mathbf{r}_i(s_i^*) = \mathbf{d}_i(s_i^*) = s_i$ *that is structurally rational for any* $\mu_i^* \in \Delta(S_{-i}^*, \mathcal{I}_i^*)$ *that agrees with* μ_i.

This result also yields a positive theoretical rationale for the use of the strategy method, provided direct play is implemented as described in Definition 5. Suppose the experimenter wishes to test whether play conforms to some solution concept that adopts structural rationality as the notion of best reply. Then, if indeed players conform to such a solution concept, the version of the strategy method proposed here will elicit their reported behavior.

COROLLARY 1: *Suppose that* $W_i = \{*\}$ *for all* $i \in N$. *Then, for all* $i \in N$ *and all* $s_i^* \in S_i^*$ *such that* $\mathbf{r}_i(s_i^*) = \mathbf{d}_i(s_i^*)$, s_i^* *is structurally rational given* μ_i^* *in the elicitation game if and only if* $\mathbf{d}_i(s_i^*)$ *is structurally rational given* μ_i *in the original game*.

Theorem 2 depends crucially on structural rationality. (Weak) sequential rationality is not sufficient, even if beliefs satisfy the agreement condition of Definition 6.

EXAMPLE 4: Consider the game in Figure 1 and assume that $W_a = \{*\}$ and $W_b = \{\{\text{In}S\}, 0.5\}$, with $I_b = J$; that is, Bob is asked to bet on Ann playing S at I, and no bet is offered to Ann. For $0 < \epsilon < 1$, the following strategies are part of a sequential equilibrium. Ann plays $(\overline{\text{Out}}, *)$ at I_a^1; Bob plays $(\bar{S}, 0.5)$ at I_b^1. If selected, Ann plays Out at information set $\phi_{\bar{t}_a,*}$ and S at information set $I_{\bar{t}_a,*}$, for all $\bar{t}_a \in S_a$; and if selected, Bob plays S at $J_{\bar{t}_b,v_b}$, for all $(\bar{t}_b, v_b) \in S_b \times W_b$. Moreover, at all $\phi_{\bar{t}_a,*}$ and $I_{\bar{t}_a,*}$, as well as at I_a^1, Ann assigns probability one to Bob having chosen reported strategy $(\bar{S}, 0.5)$; at I_b^1, Bob expects Ann to have chosen $(\overline{\text{Out}})$, and at each $J_{\bar{t}_b,v_b}$, he assigns probability one to Ann having chosen reported strategy $\overline{\text{In}S}$.

The key is that Bob must bet at the beginning of the game, where sequential rationality[10] only requires that he maximize his ex ante expected payoff. In equilibrium, Bob expects Ann to choose $\overline{\text{Out}}$, so that the bet is called off; hence, he is indifferent between his betting choices.

To reconcile Theorem 2 and Example 4 with the generic equivalence results described in Section 5.2, notice that elicitation games feature numerous relevant ties *by construction*. Consider Bob at I_b^1 in Figure 4. If Ann reports strategy $\overline{\text{Out}}$ at I, then for a fixed report $\bar{s}_b$ of Bob, both actions $(\bar{s}_b, \{\text{In}S\})$ and $(\bar{s}_b, 0.5)$ yield the same payoff, namely $2 + \epsilon$. This is a relevant tie.

[10]Here, the distinction between weak and full sequential rationality is immaterial. The profile described in the example is part of a sequential equilibrium.

6. DISCUSSION

6.1. *Incomplete-Information Games*

To accommodate incomplete information, fix a dynamic game with N players, strategy sets S_i, terminal histories Z, and information sets $\mathcal{I}_i$ for each $i \in N$. Consider finite sets Θ_i of "types" for each $i \in N$, and a set Θ_0 that captures residual uncertainty not reflected in players' types. Player i's payoff function is a map $u_i : Z \times \Theta \to \mathbb{R}$, where $\Theta = \Theta_0 \times \prod_{j \in N} \Theta_j$. The conditional beliefs of player i's type θ_i can then be represented via a CCPS $\mu_{\theta_i} \in \Delta(S_{-i} \times \Theta)^{\{\phi\} \cup \mathcal{I}_i}$; now a perturbation is a sequence $(p^k)_{k \geq 1} \subset \Delta(S_{-i} \times \Theta)$ such that $p^k(S_{-i}(I) \times \Theta) > 0$ and $p^k(S_{-i}(I) \times \Theta) \to \mu_{\theta_i}(\cdot|I)$ for all $I \in \{\phi\} \cup \mathcal{I}_i$. Definitions 1, 2, and 3 can be applied to each type $\theta_i \in \Theta_i$ separately; Theorems 1, 2, and 3 have straightforward extensions. If the sets Θ_i are infinite, the alternative characterization of structural rationality in Siniscalchi (2022) is a more convenient starting point, but Theorems 1 and 2 still hold.

6.2. *Higher-Order Beliefs*

The proposed approach can also be adapted to elicit higher-order beliefs. Consider a two-player game for simplicity. The analyst first elicits Ann's first-order beliefs about Bob's strategies, as in Section 5.3. She then elicits Bob's second-order beliefs by offering him side bets on both Ann's strategies *and* on her first-order beliefs. To formalize this, one follows §6.1, taking Θ_i to be the set of all CCPSs for each player i. The incomplete-information extension of Theorem 2 ensures that second-order beliefs can be elicited in an incentive-compatible way. The argument extends to beliefs of higher orders.

6.3. *Elicitation: Modified or Perturbed Games*

In the equilibrium (Out, (S, S)) of the game of Figure 1, Ann's initial move prevents J from being reached. One might consider modifying the game so that J *is* actually reached, perhaps with small probability, regardless of Ann's initial move. However, such modifications may have a significant impact on players' strategic reasoning and behavior and, therefore, on elicited beliefs. For instance, in the game of Figure 1, *forward-induction* reasoning selects the equilibrium (In, (B, B)) (cf., e.g., Van Damme (1989)). Thus, if Ann follows the logic of forward induction, she should expect Bob to play B. However, suppose action Out is removed. Then the game reduces to the simultaneous-move Battle of the Sexes, in which forward induction has no bite. Ann may well expect Bob to play B in the game of Figure 1, and S in the game with Out removed. Thus, Ann's beliefs elicited in the latter game may differ from her actual beliefs in the former. Similar conclusions hold if one causes Ann to play In with positive probability when she chooses Out. Analogous arguments apply to backward-induction reasoning; see, for example, Ben-Porath (1997, Example 3.2 and page 36).

By way of contrast, the elicitation approach in Section 5.3 only modifies the game in ways that, as per part (i) of Theorem 2, are inessential under structural rationality.

6.4. *Caution, Elicitation, and Triviality*

Consider the games in Figure 5(a).[11] Ann has a single move available at I in Figure 5(a).

[11] I thank a referee for providing this example, which motivated the discussion in this subsection.

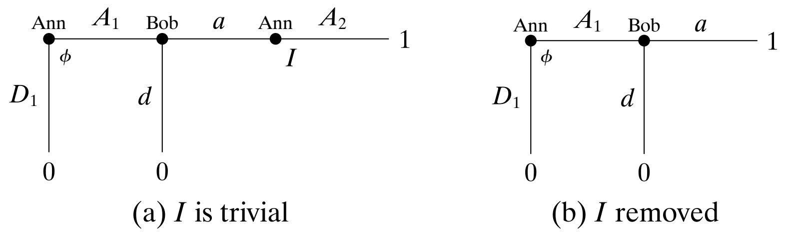

FIGURE 5.—A trivial information set; Ann's payoffs shown.

From the perspective of weak sequential rationality, such an information set can be disregarded. However, if Ann assigns prior probability 1 to d, A_1A_2 is the only structurally rational strategy for her in in Figure 5(a), whereas both D_1 and A_1A_2 are structurally rational in Figure 5(b). The reason is that Ann has different conditioning events in the two games in Figure 5.

To avoid this, one can replace $\mathcal{I}_i$ with the collection $\mathcal{I}_i^{\text{nt}} = \{I \in \mathcal{I}_i : |A(I)| \geq 2\}$ of "non-trivial" information sets in Definitions 1, 2, 4, and 6; all of the results in this paper continue to hold (except that, naturally, beliefs at trivial information sets can no longer be elicited). In fact, "trivial" information sets are only used to model the experimenter's mechanical implementation of subjects' reported strategies in Definition 5.

6.5. *Equilibrium and Structurally Rational Strategies*

To illustrate how structural rationality can be incorporated into solution concepts, consider Govindan and Wilson's (2009) reformulation of sequential equilibrium. A *behavioral strategy* for player i is an array $\beta = (\beta_i(I))_{I\in\mathcal{I}_i} \in \Delta(A)^{\mathcal{I}_i}$ such that $\beta_i(I)(A(I)) = 1$ for all $I \in \mathcal{I}_i$. As usual, each behavioral strategy β_i induces a mixed strategy $\sigma_i \in \Delta(S_i)$; $\bigotimes_{j\neq i}\sigma_j$ denotes the product measure with marginals σ_j, for $j \neq i$. Then a *sequential equilibrium* is a profile $(\beta_i, \mu_i)_{i\in N}$ where each β_i is a behavioral strategy for i, $\mu_i = (\mu_i(\cdot|I))_{I\in\mathcal{I}} \in \Delta(S_{-i})^{\{\phi\}\cup\mathcal{I}_i}$, and the following two conditions hold:

(i) There is a sequence of strictly positive behavioral strategy profiles $(\beta_i^k)_{i\in N,k\geq 1}$ and a sequence of strictly positive mixed strategy profiles $(\sigma_i^k)_{i\in N,k\geq 1}$ such that, for every i, each σ_i^k is derived from β_i^k, $\beta_i^k \to \beta_i$, and $(\bigotimes_{j\neq i} p_j^k)(\cdot|S_{-i}(I)) \to \mu_i(\cdot|I)$ for each $I \in \mathcal{I}_i$.

(ii) For every i and $I \in \mathcal{I}_i$, if $\beta_i(I)(a) > 0$ then there exists $s_i \in S_i(I)$ such that $s_i(I) = a$ and $s_i \in \arg\max_{t_i\in S_i(I)} U_i(t_i, \mu_i(\cdot|I))$.

By condition (i), each μ_i is a CCPS, generated by a specific type of perturbation.

To obtain a corresponding notion of "*structural equilibrium*," replace (ii) above with:

(ii') For every i and $I \in \mathcal{I}_i$, if $\beta_i(I)(a) > 0$, then there exists $s_i \in S_i(I)$ such that $s_i(I) = a$ and a perturbation $(p^k)_{k\geq 1}$ of μ_i such that $s_i \in \arg\max_{t_i\in S_i(I)} U_i(t_i, p^k(\cdot|S_{-i}(I))$ for all $k \geq 1$.

Refer to the companion paper Siniscalchi (in preparation) for further analysis.

In addition, there is a straightforward relationship with solution concepts based on "trembles:" *only structurally rational strategies are played in a trembling-hand perfect equilibrium* (Selten (1975)). In the notation of this paper, a (strategic-form) *(trembling-hand) perfect equilibrium* is a profile $\sigma \in \prod_{i\in I}\Delta(S_i)$ such that, for every $i \in N$, there exists a sequence $(\sigma_i^k)_{k\geq 1}$ such that $\sigma_i^k \to \sigma_i$ and every $s_i \in \operatorname{supp}\sigma_i$ is a best reply to each product measure $p_{-i}^k \equiv \bigotimes_{j\neq i}\sigma_j^k$, $k \geq 1$. Each sequence $(p_{-i}^k)_{k\geq 1}$ defines a CCPS $\mu_{-i} \in \Delta(S_{-i}, \mathcal{I}_i)$

(possibly considering subsequences), and by Definition 3, every $s_i \in \operatorname{supp} \sigma_i$ is structurally rational given μ_{-i}.

APPENDIX A: DYNAMIC GAMES

Fix a dynamic game $(N, A, Z, P, (\mathcal{I}_i, u_i)_{i\in N})$ as defined in Section 2. Let H be the set of all (terminal and nonterminal) histories, as defined therein.[12]

Let $h = (a_1, \ldots, a_K) \in H$. For all $k = 0, \ldots, K-1$, $h' \equiv (a_1, \ldots, a_k)$ is a *prefix* of h, written $h' < h$. The case $k = 0$ corresponds to $h' = \phi$, which is a prefix of every history. I sometimes write $h' \leq h$ to mean that either $h' = h$ or h' is a prefix of h.

Perfect recall is formalized per Definition 203.3 in Osborne and Rubinstein (1994). For every $h \in P^{-1}(i)$, let $X_i(h)$ denote i's *experience* along the history h; if $h = (a_1, \ldots, a_L)$, let $\ell_1, \ldots, \ell_K$ be the set of indices $\ell \in \{1, \ldots, L-1\}$ such that $P((a_1, \ldots, a_{\ell-1})) = i$, and $I_1, \ldots, I_K$ be such that $(a_1, \ldots, a_{\ell_k-1}) \in I_k$ for $k = 1, \ldots, K$; then $X_i(h) = (I_1, a_{\ell_1}, \ldots, I_k, a_{\ell_k})$. Perfect recall requires that, if $h, h' \in I \in \mathcal{I}_i$, then $X_i(h) = X_i(h')$. One immediate implication (used in the proof of Remark 1) is that, if $h < h'$, then h and h' cannot be elements of the same information set.

The *terminal history map* $\zeta : S \to Z$ associates with each strategy profile s the terminal history it induces; that is, $\zeta(s) = z$ iff $s \in S(z)$.

REMARK 1: Let $h = (a_1, \ldots, a_K) \in H$. Then, for every $i \in N$, $s_i \in S_i(h) \equiv \operatorname{proj}_{S_i} S(h)$ if and only if, for every $k = 1, \ldots, K$, if $P((a_1, \ldots, a_{k-1})) = i$ and $I \in \mathcal{I}_i$ is the unique information set such that $(a_1, \ldots, a_{k-1}) \in I$, then $s_i(I) = a_k$. In particular, $S(h) = \prod_{i\in N} S_i(h)$.

PROOF: Suppose that $s_i \in S_i(h)$, so by definition there is $s_{-i} \in S_{-i}$ such that $(s_i, s_{-i}) \in S(h)$. Since $\mathcal{I}_j$ is a is a partition of $P^{-1}(\{j\}) \subseteq H \setminus Z$ for all $j \in N$, for every $k = 1, \ldots, K$, if $i = P((a_1, \ldots, a_{k-1}))$, then $(a_1, \ldots, a_{k-1}) \in I \in \mathcal{I}_j$ implies that $j = i$. Hence, $(s_i, s_{-i}) \in S(h)$ implies $s_i(I) = a_k$.

Conversely, suppose that, for some $s_i \in S_i$, and for all k with $P((a_1, \ldots, a_{k-1})) = i$, $s_i(I) = a_k$, where $(a_1, \ldots, a_{k-1}) \in I \in \mathcal{I}_i$. Define $s_{-i} \in S_{-i}$ as follows: for every $j \neq i$ and all $J \in \mathcal{I}_j$, if $(a_1, \ldots, a_{k-1}) \in J$ for some k, then $s_j(J) = a_k$; otherwise $s_j(J)$ is an arbitrary element of $A(J)$. By perfect recall, there is at most one k such that $(a_1, \ldots, a_{k-1}) \in J$, so this definition is well posed. Furthermore, by construction the profile (s_i, s_{-i}) is such that $P((a_1, \ldots, a_{k-1})) = j$ and $(a_1, \ldots, a_{k-1}) \in J \in \mathcal{I}_j$ imply $s_j(J) = a_k$, regardless of whether $j = i$ or $j \neq i$. Hence, $(s_i, s_{-i}) \in S(h)$, so $s_i \in \operatorname{proj}_{S_i} S(h) = S_i(h)$. *Q.E.D.*

REMARK 2: For all $i \in N$ and $I \in \mathcal{I}_i$, $S(I) = S_i(I) \times S_{-i}(I)$.[13]

PROOF: $s_i \in S_i(I)$ implies that there is $t_{-i} \in S_{-i}(I)$ with $(s_i, t_{-i}) \in S(I)$. Similarly, $s_{-i} \in S_{-i}(I)$ implies that there is $t_i \in S_i$ with $(t_i, s_{-i}) \in S(I)$. Let $h', h'' \in I$ be such that $(s_i, t_{-i}) \in S(h')$ and $(t_i, s_{-i}) \in S(h'')$. By perfect recall, $X_i(h') = X_i(h'') \equiv (I_1, a_1, \ldots, I_K, a_K)$. Let $\bar{h}'' < h''$ be such that $P(\bar{h}'') = i$. By the definition of $X_i(\cdot)$, there is k such that $\bar{h}'' \in I_k$. Then there must be $\bar{h}' < h'$ such that $\bar{h}' \in I_k$ as well, and $s_i(I_k) = a_k = t_i(I_k)$; otherwise $X_i(h') \neq X_i(h'')$. By Remark 1, this implies that $(s_i, s_{-i}) \in S(h'')$, and so $(s_i, s_{-i}) \in S(I)$, as claimed. *Q.E.D.*

[12] Osborne and Rubinstein (1994) take as primitive a set H of histories, closed under the "subhistory" (prefix) relation, and define Z as the set of histories that are no proper prefix of any other history. The approach taken here starts from Z and derives H; it is more convenient in Definition 5, but equivalent.

[13] This result is known, but I have been unable to find a published proof.

APPENDIX B: PROOFS OF THE MAIN RESULTS

B.1. *Proof of Proposition 1*

If $s_i \in S_i$ is structurally rational for μ, there is a perturbation $(p^k)_{k\geq 1}$ of μ such that $U_i(s_i, p^k) \geq U_i(t_i, p^k)$ for all $t_i \in S_i$ and all $k \geq 1$. Hence, for all $\sigma_i \in \Delta(S_i)$, $U_i(s_i, p^k) \geq U_i(\sigma_i, p^k)$ for all $k \geq 1$, so no $\sigma_i \in \Delta(S_i)$ satisfies $U_i(\sigma_i, p^k) > U_i(s_i, p^k)$ eventually for *all* perturbations $(p^k)_{k\geq 1}$ of μ.

Now suppose s_i is not structurally rational for μ. Denote by $v(s_{-i})$ the s_{-i}-th coordinate of $v \in \mathbb{R}^{S_{-i}}$. For $t_i \in S_i$, $I \in \mathcal{I}_i \cup \{\phi\}$, $s_{-i} \in S_{-i}(I)$, and $\epsilon > 0$, define a^I, a^{t_i}, $a_\epsilon^{I,s_{-i},+}$, $a_\epsilon^{I,s_{-i},-} \in \mathbb{R}^{S_{-i}}$ by

- $a^I(t_{-i}) = -1$ if $t_{-i} \in S_{-i}(I)$, and $a^I(t_{-i}) = 0$ for $t_{-i} \notin S_{-i}(I)$;
- $a^{t_i}(t_{-i}) = U_i(t_i, t_{-i}) - U_i(s_i, t_{-i})$ for all $t_{-i} \in S_{-i}$;
- $a_\epsilon^{I,s_{-i},+}(t_{-i}) = -[\mu(\{s_{-i}\}|I) + \epsilon]$ for $t_{-i} \in S_{-i}(I) \setminus \{s_{-i}\}$, $a_\epsilon^{I,s_{-i},+}(s_{-i}) = 1 - [\mu(\{s_{-i}\}|I) + \epsilon]$, and $a_\epsilon^{I,s_{-i},+}(t_{-i}) = 0$ for $t_{-i} \notin S_{-i}(I)$;
- $a_\epsilon^{I,s_{-i},-}(t_{-i}) = [\mu(\{s_{-i}\}|I) - \epsilon]$ for $t_{-i} \in S_{-i}(I) \setminus \{s_{-i}\}$, $a_\epsilon^{I,s_{-i},-}(s_{-i}) = -1 + [\mu(\{s_{-i}\}|I) - \epsilon]$, and $a_\epsilon^{I,s_{-i},-}(t_{-i}) = 0$ for $t_{-i} \notin S_{-i}(I)$.

Let $m \in \mathbb{R}_+^{S_{-i}}$ and consider the following system of linear inequalities:

$$a^I \cdot m \leq -1 \quad \forall I \in \mathcal{I}_i \cup \{\phi\}, \tag{12}$$

$$a^{t_i} \cdot m \leq 0 \quad \forall t_i \in S_i, \tag{13}$$

$$a_\epsilon^{I,s_{-i},+} \cdot m \leq 0 \quad \forall I \in \mathcal{I}_i \cup \{\phi\}, s_{-i} \in S_{-i}(I), \tag{14}$$

$$a_\epsilon^{I,s_{-i},-} \cdot m \leq 0 \quad \forall I \in \mathcal{I}_i \cup \{\phi\}, s_{-i} \in S_{-i}(I). \tag{15}$$

By contradiction, suppose the system defined by equations (12)–(15) has a solution for every $\epsilon > 0$. For each $k \geq 1$, let m^k be a solution for $\epsilon = \frac{1}{k}$. From equation (12) and the definition of a^I, $M^k(I) \equiv \sum_{s_{-i}\in S_{-i}(I)} m^k(s_{-i}) \geq 1$ for all $I \in \mathcal{I}_i \cup \{\phi\}$; in particular, $M^k(\phi) > 0$, and one can define $p^k \in \Delta(S_{-i})$ by letting $p^k(\{s_{-i}\}) = m^k(s_{-i})/M^k(\phi)$ for all $s_{-i} \in S_{-i}$. Then, for all $I \in \mathcal{I}_i \cup \{\phi\}$, $p^k(S_{-i}(I)) = \sum_{s_{-i}\in S_{-i}(I)} m^k(s_{-i})/M^k(\phi) = M^k(I)/M^k(\phi) \geq 1/M^k(\phi) > 0$ because $M^k(I) \geq 1$. Now Equations (14) and (15) and the definition of $a_\epsilon^{I,s_{-i},+}$, $a_\epsilon^{I,s_{-i},-}$, and $M^k(I)$ imply that

$$m^k(s_{-i}) - \mu\big(\{s_{-i}\}|I\big)M^k(I) \leq \frac{1}{k}M^k(I) \quad \text{and} \quad -m^k(s_{-i}) + \mu\big(\{s_{-i}\}|I\big)M^k(I) \leq \frac{1}{k}M^k(I),$$

that is, $|m^k(s_{-i}) - \mu(\{s_{-i}\}|I)M^k(I)| \leq \frac{1}{k}M^k(I)$, for every $I \in \mathcal{I}_i \cup \{\phi\}$ and $s_{-i} \in S_{-i}(I)$. Dividing by $M^k(I)$, since $m^k(s_{-i})/M^k(I) = \frac{m^k(s_{-i})/M^k(\phi)}{M^k(I)/M^k(\phi)} = \frac{p^k(\{s_{-i}\})}{p^k(S_{-i}(I))} = p^k(\{s_{-i}\}|S_{-i}(I))$, one has $|p^k(\{s_{-i}\}|S_{-i}(I)) - \mu(\{s_{-i}\}|I)| < \frac{1}{k}$, so $p^k(\{s_{-i}\}|S_{-i}(I)) \to \mu(\{s_{-i}\}|I)$. Hence, $(p^k)_{k\geq 1}$ is a perturbation of μ. Finally, for every $t_i \in S_i$, by equation (13) and the definition of a^{t_i}, $\sum_{s_{-i}}[U_i(t_i, s_{-i}) - U_i(s_i, s_{-i})]m(s_{-i}) \leq 0$, so dividing by $M^k(\phi)$, $\sum_{s_{-i}}[U_i(t_i, s_{-I}]) - U_i(s_i, s_{-i})]p^k(\{s_{-i}\}) \leq 0$, that is, $U_i(s_i, p^k) \geq U_i(t_i, p^k)$. Since this holds for every k and t_i, s_i is structurally rational given μ, contradiction.

Thus, for some $\epsilon > 0$, the system defined by equations (12)–(15) has no solution. Then, by Theorem 22.1 in Rockafellar (1970) (a version of the theorem of the alternative), there exist $\lambda^I \geq 0$ for every $I \in \mathcal{I}_i \cup \{\phi\}$, $\lambda^{t_i} \geq 0$ for every $t_i \in S_i$, $\lambda^{I,s_{-i},+} \geq 0$ for every $I \in \mathcal{I}_i \cup \{\phi\}$

and $s_{-i} \in S_{-i}(I)$, and $\lambda^{I,s_{-i},-} \geq 0$ for every $I \in \mathcal{I}_i \cup \{\phi\}$ and $s_{-i} \in S_{-i}(I)$, such that

$$\sum_{I \in \mathcal{I}_i \cup \{\phi\}} \lambda^I a^I + \sum_{t_i \in S_i} \lambda^{t_i} a^{t_i} + \sum_{I \in \mathcal{I}_i \cup \{\phi\}, s_{-i} \in S_{-i}(I)} \lambda^{I,s_{-i},+} a^{I,s_{-i},+} + \sum_{I \in \mathcal{I}_i \cup \{\phi\}, s_{-i} \in S_{-i}(I)} \lambda^{I,s_{-i},-} a^{I,s_{-i},-} = \mathbf{0}, \tag{16}$$

where $\mathbf{0}$ is the zero vector in $\mathbb{R}^{S_{-i}}$, and furthermore,

$$\sum_{I \in \mathcal{I}_i \cup \{\phi\}} \lambda^I \cdot (-1) + \sum_{t_i \in S_i} \lambda^{t_i} \cdot 0 + \sum_{I \in \mathcal{I}_i \cup \{\phi\}, s_{-i} \in S_{-i}(I)} \lambda^{I,s_{-i},+} \cdot 0 + \sum_{I \in \mathcal{I}_i \cup \{\phi\}, s_{-i} \in S_{-i}(I)} \lambda^{I,s_{-i},-} \cdot 0 < 0. \tag{17}$$

I show that $\sum_{t_i} \lambda^{t_i}[U(t_i, p^k) - U_i(s_i, p^k)] > 0$ eventually for all perturbations $(p^k)_{k \geq 1}$ of μ. This also implies that $\Lambda \equiv \sum_{t_i} \lambda^{t_i} > 0$, so to complete the proof one can let $\sigma_i = (\frac{\lambda^{t_i}}{\Lambda})_{t_i \in S_i}$.

Fix one such perturbation $(p^k)_{k \geq 1}$. Then $p^k(S_{-i}(I)) > 0$ for all $I \in \mathcal{I}$, and $p^k(\{s_{-i}\}| S_{-i}(I)) \to \mu(\{s_{-i}\}|I)$ for all $I \in \mathcal{I}_i \cup \{\phi\}$ and $s_{-i} \in S_{-i}(I)$. Thus, for some $K \geq 1$, $k \geq K$ implies $|p^k(\{s_{-i}\}|S_{-i}(I)) - \mu(\{s_{-i}|I)| \leq \epsilon$. Let $p^k_{\min} = \min_{I \in \mathcal{I}_i \cup \{\phi\}} p^k(S_{-i}(I))$. Then $p^k_{\min} > 0$ and $p^k(S_{-i}(I)) \geq p^k_{\min}$ for all $I \in \mathcal{I}_i \cup \{\phi\}$. Abusing notation, let $a \cdot p^k \equiv \sum_{t_{-i}} a(t_{-i}) \cdot p^k(\{t_{-i}\})$ for every $a \in \mathbb{R}^{S_{-i}}$. Then $a^I \cdot p^k = -p^k(S_{-i}(I)) \leq -p^k_{\min} < 0$ for all $I \in \mathcal{I}_i \cup \{\phi\}$. Furthermore, $|p^k(\{s_{-i}\}|S_{-i}(I)) - \mu(\{s_{-i}|I)| \leq \epsilon$ iff $p^k(\{s_{-i}\}|S_{-i}(I)) - \mu(\{s_{-i}\}|I) \leq \epsilon$ and $-p^k(\{s_{-i}\}|S_{-i}(I)) + \mu(\{s_{-i}\}|I) \leq \epsilon$, that is, multiplying by $p^k(S_{-i}(I)) > 0$, iff $p^k(\{s_{-i}\}) - \mu(\{s_{-i}\}|I)p^k(S_{-i}(I)) \leq \epsilon p^k(S_{-i}(I))$ and $-p^k(\{s_{-i}\}) + \mu(\{s_{-i}\}|I)p^k(S_{-i}(I)) \leq \epsilon p^k(S_{-i}(I))$, or $p^k(\{s_{-i}\}) - [\mu(\{s_{-i}\}|I) + \epsilon]p^k(S_{-i}(I)) \leq 0$ and $-p^k(\{s_{-i}\}) + [\mu(\{s_{-i}\}|I) - \epsilon]p^k(S_{-i}(I)) \leq 0$; that is, $a_\epsilon^{I,s_{-i},+} \cdot p^k \leq 0$ and $a_\epsilon^{I,s_{-i},-} \cdot p^k \leq 0$ for all $I \in \mathcal{I}_i \cup \{\phi\}$ and $s_{-i} \in S_{-i}(I)$.

Now for each $t_{-i} \in S_{-i}$, taking the dot product of each side of equation (16) with p^k yields

$$\sum_{I \in \mathcal{I}_i \cup \{\phi\}} \lambda^I a^I \cdot p^k + \sum_{t_i \in S_i} \lambda^{t_i} a^{t_i} \cdot p^k + \sum_{I \in \mathcal{I}_i \cup \{\phi\}, s_{-i} \in S_{-i}(I)} \lambda^{I,s_{-i},+} a_\epsilon^{I,s_{-i},+} \cdot p^k + \sum_{I \in \mathcal{I}_i \cup \{\phi\}, s_{-i} \in S_{-i}(I)} \lambda^{I,s_{-i},-} a_\epsilon^{I,s_{-i},-} \cdot p^k = 0.$$

Since $\lambda^{I,s_{-i},+}, \lambda^{I,s_{-i},-} \geq 0$, $a_\epsilon^{I,s_{-i},+} \cdot p^k \leq 0$, and $a_\epsilon^{I,s_{-i},-} \cdot p^k \leq 0$ for all $I \in \mathcal{I}_i \cup \{\phi\}$ and $s_{-i} \in S_{-i}(I)$, the third and fourth summations are nonpositive. Also, for every $I \in \mathcal{I}_i \cup \{\phi\}$, $a^I \cdot p^k < 0$, and by equation (17), $\lambda^I > 0$ for at least one $I \in \mathcal{I}_i \cup \{\phi\}$: thus, the first summation is strictly negative. Hence, the second summation is strictly positive. From the definition of a^{t_i} for $t_i \in S_i$, $\sum_{t_i \in S_i} \lambda^{t_i}[U_i(t_i, p^k) - U_i(s_i, p^k)] = \sum_{t_i \in S_i} \lambda^{t_i} \sum_{t_{-i} \in S_{-i}}[U_i(t_i, t_{-i}) - U_i(s_i, t_{-i})]p^k(\{t_{-i}\}) > 0$.

B.2. *Proof of Theorem 1*

Suppose that $s_i \in S_i$ is structurally rational given μ. Fix $I \in \mathcal{I}_i$ with $s_i \in S_i(I)$ and $r_i \in S_i(I)$. By strategic independence (cf. Section 2), there is $t_i \in S_i$ such that $U_i(t_i, s_{-i}) = U_i(r_i, s_{-i})$ for $s_{-i} \in S_{-i}(I)$, and $U_i(t_i, s_{-i}) = U_i(s_i, s_{-i})$ for $s_{-i} \notin S_{-i}(I)$. By Definition 3, there is a perturbation (p^k) of μ such that $U_i(s_i, p^k) \geq U_i(t_i', p^k)$ for all $t_i' \in S_i$. In particular, for $t'_i = t_i$,

$$\begin{aligned} U_i\big(s_i, p^k\big(\cdot|S_{-i}(I)\big)\big) &= \sum_{s_{-i}\in S_{-i}(I)} U_i(s_i, s_{-i})\, p^k\big(\{s_{-i}\}|S_{-i}(I)\big) \\ &= \frac{1}{p^k\big(S_{-i}(I)\big)} \cdot \sum_{s_{-i}\in S_{-i}(I)} U_i(s_i, s_{-i})\, p^k\big(\{s_{-i}\}\big) \\ &= \frac{1}{p^k\big(S_{-i}(I)\big)}\Bigg[\sum_{s_{-i}\in S_{-i}} U_i(s_i, s_{-i})\, p^k\big(\{s_{-i}\}\big) - \sum_{s_{-i}\notin S_{-i}(I)} U_i(s_i, s_{-i})\, p^k\big(\{s_{-i}\}\big)\Bigg] \\ &\geq \frac{1}{p^k\big(S_{-i}(I)\big)}\Bigg[\sum_{s_{-i}\in S_{-i}} U_i(t_i, s_{-i})\, p^k\big(\{s_{-i}\}\big) - \sum_{s_{-i}\notin S_{-i}(I)} U_i(s_i, s_{-i})\, p^k\big(\{s_{-i}\}\big)\Bigg] \\ &= \frac{1}{p^k\big(S_{-i}(I)\big)} \sum_{s_{-i}\in S_{-i}(I)} U_i(r_i, s_{-i})\, p^k\big(\{s_{-i}\}\big) = U_i\big(r_i, p^k\big(\cdot|S_{-i}(I)\big)\big). \end{aligned}$$

The second equality follows from the definition of conditional probability and the fact that, by Definition 1, $p^k(S_{-i}(I)) > 0$. The inequality follows from the choice of the perturbation $(p^k)_{k\geq 1}$. The fourth equality follows from the definition of t_i. Since $p^k(\cdot|S_{-i}(I)) \to \mu(\cdot|I)$ by Definition 1, it follows that $U_i(s_i, \mu(\cdot|I)) \geq U_i(r_i, \mu(\cdot|I))$.

B.3. *Generic Equivalence of Structural and Sequential Rationality*

A *relevant tie* for player i is a tuple (I, s_i, t_i, t_{-i}) such that $I \in \mathcal{I}_i \cup \{\phi\}$, $s_i, t_i \in S_i(I)$, $t_{-i} \in S_{-i}(I)$, $\zeta(s_i, t_{-i}) \neq \zeta(t_i, t_{-i})$, and $U_i(s_i, t_{-i}) = U_i(t_i, t_{-i})$. That is, starting from I, if coplayers move according to t_{-i}, then i's strategies s_i and t_i reach distinct terminal histories, but yield the same payoff. A *nontrivial redundance* for player i is a tuple $(I, s_i, \sigma_i, t_{-i}, t'_{-i})$ such that $I \in \mathcal{I}_i \cup \{\phi\}$, $s_i \in S_i(I)$, $\sigma_i(S_i(I) \setminus \{s_i\}) = 1$, $t_{-i}, t'_{-i} \in S_{-i}(I)$, $U_i(s_i, s_{-i}) = U_i(\sigma_i, s_{-i})$ for $s_{-i} \in \{t_{-i}, t'_{-i}\}$, and $\zeta(t_i, t_{-i}) \neq \zeta(t_i, t'_{-i})$ for some $t_i \in \operatorname{supp} \sigma_i$. That is, the payoff that s_i yields given t_{-i} and t'_{-i} is a nontrivial[14] convex combination of payoffs of other strategies of i at I.

In the game in Figure 1, there are no relevant ties or nontrivial redundances for Ann. On the other hand, in Figure 2, if $x = 2$, there is a relevant tie at the initial node.

THEOREM 3: *Fix $i \in N$ and $\mu \in \Delta(S_{-i}, \mathcal{I}_i)$. If $s_i \in S_i$ is weakly sequentially rational given μ, and there is no relevant tie or nontrivial redundance for i, then s_i is structurally rational given μ.*

PROOF: Assume that $s_i \in S_i$ is weakly sequentially rational given μ, and that the game has no relevant ties or nontrivial redundancies for i. We show that, for every $\sigma_i \in \Delta(S_i)$, there is a perturbation $(\tilde{p}^k)_{k\geq 1}$ of μ for which $U_i(s_i, \tilde{p}^k) \geq U_i(\sigma_i, \tilde{p}^k)$ for all k; by Proposition 1, this implies that s_i is structurally rational for μ. Thus, fix $\sigma_i \in \Delta(S_i)$. If $\sigma_i(\{s_i\}) < 1$, then for every $p \in \Delta(S_{-i})$, $U_i(s_i, p) \geq U_i(\sigma_i, p) = \sigma_i(\{s_i\})U_i(s_i, p) + [1 -$

[14]If $\zeta(t_i, t_{-i}) = \zeta(t_i, t'_{-i})$ for all $t_i \in \operatorname{supp} \sigma_i$, then either there is a relevant tie, or for small payoff perturbations, one can correspondingly perturb σ_i so that the condition "$U_i(s_i, s_{-i}) = U_i(\sigma_i, s_{-i})$ for $s_{-i} \in \{t_{-i}, t'_{-i}\}$" holds.

$\sigma_i(\{s_i\})]U_i(\sigma_i(\cdot|S_i\setminus\{s_i\}),p)$ holds iff $U_i(s_i,p)\geq U_i(\sigma_i(\cdot|S_i\setminus\{s_i\}),p)$. Thus, it is enough to prove the result for σ_i with $\sigma_i(\{s_i\})=0$.

For every $s_{-i}\in S_{-i}$, let $h(s_{-i})\in H$ be the longest history h such that $h\leq\zeta(s_i,s_{-i})$ and $h\leq\zeta(t_i,s_{-i})$ for all $t_i\in\operatorname{supp}\sigma_i$. If $h(s_{-i})=\zeta(s_i,s_{-i})$, then also $h(s_i)=\zeta(t_i,s_{-i})$ for all $t_i\in\operatorname{supp}\sigma_i$, because terminal histories are not ranked by the prefix relation; conversely, for the same reason, if $h(s_{-i})=\zeta(t_i,s_{-i})$ for some $t_i\in\operatorname{supp}\sigma_i$, then in fact $h(s_{-i})=\zeta(t_i',s_{-i})$ for all $t_i'\in\operatorname{supp}\sigma_i$, and $h(s_{-i})=\zeta(s_i,s_{-i})$. Furthermore, if $h(s_{-i})\in H\setminus Z$, then $P(h)=i$; by contradiction, if $P(h(s_{-i}))=j\neq i$, then $h(s_{-i})\in J$ for some $J\in\mathcal{I}_j$, so that $(h(s_{-i}),s_j(J))\leq\zeta(s_i,s_{-i})$ and $(h(s_{-i}),s_j(J))\leq\zeta(t_i,s_{-i})$ for all $t_i\in\operatorname{supp}\sigma_i$, which contradicts the definition of $h(s_{-i})$. Hence, either $h(s_{-i})\in Z$, in which case $h(s_{-i})=\zeta(s_i,s_{-i})=\zeta(t_i,s_{-i})$ for all $t_i\in\operatorname{supp}\sigma_i$, or else $h(s_{-i})\in I$ for some $I\in\mathcal{I}_i$; in the latter case, denote the unique element of $\mathcal{I}_i$ containing $h(s_{-i})$ by $I(s_{-i})$; then, by the definition of $h(s_{-i})$, $s_i(I(s_{-i}))\neq t_i(I(s_{-i}))$ for at least one $t_i\in\operatorname{supp}\sigma_i$, for otherwise $a=s_i(I(s_{-i}))$ would satisfy $a=t_i(I(s_{-i}))$ for all $t_i\in\operatorname{supp}\sigma_i$, $(h(s_{-i}),a)\leq\zeta(s_i,s_{-i})$, and $(h(s_{-i}),a)\leq\zeta(t_i,s_{-i})$ for all $t_i\in\operatorname{supp}\sigma_i$, contradiction. Furthermore, $\sigma_i(S_i(I(s_{-i})))=1$, because every $t_i\in\operatorname{supp}\sigma_i$ satisfies $t_i\in S_i(h(s_{-i}))\subseteq S_i(I(s_{-i}))$. Since $\sigma_i(\{s_i\})=0$, $\sigma_i(S_i(I(s_{-i}))\setminus\{s_i\})=1$.

Now fix $s_{-i},t_{-i}\in S_{-i}$. I claim that either $S_{-i}(I(s_{-i}))\cap S_{-i}(I(t_{-i}))=\emptyset$, or $S_{-i}(I(s_{-i}))=S_{-i}(I(t_{-i}))$. Suppose that there is $r_{-i}\in S_{-i}(I(s_{-i})\cap S_{-i}(I(t_{-i}))$. Since $s_i\in S_i(I(s_{-i}))\cap S_i(I(t_{-i}))$, by perfect recall, there are $h\in I(s_{-i})$ with $h<\zeta(s_i,r_{-i})$ and $h'\in I(t_{-i})$ with $h'<\zeta(s_i,r_{-i})$. Since h and h' are prefixes of the same terminal history, either they coincide, or they are ordered by precedence. If $h<h'$, then $I(s_{-i})$ is in i's experience at h', and hence, by perfect recall, at $h(t_{-i})$. Hence, there must be $h''<h(t_{-i})$ such that $h''\in I(s_{-i})$. Perfect recall also implies that, for some $a\in A$, $(h,a)\leq h'$ and $(h'',a)\leq h(t_{-i})$: hence, all strategies in $S_i(I(t_{-i}))$ must play a at $I(s_{-i})$, so in particular $s_i(I(s_{-i}))=a=t_i(I(s_{-i}))$ for all $t_i\in\operatorname{supp}\sigma_i$, which contradicts the fact that, as was shown above, $s_i(I(s_{-i}))\neq t_i'(I(s_{-i}))$ for at least some $t_i'\in\operatorname{supp}\sigma_i$. Similarly, it cannot be that $h'<h$. Thus, $h=h'$, and so $h=h'\in I(s_{-i})\cap I(t_{-i})$. Since $\mathcal{I}_i$ partitions $P^{-1}(\{i\})$, $I(s_{-i})=I(t_{-i})$. Therefore, writing $S^0_{-i}=\{s_{-i}:h(s_{-i})\in Z\}\}$ and arbitrarily enumerating the collection $\{I(s_{-i}):s_{-i}\in S_{-i}\}$ as $I_1,\ldots,I_L$, $\{S^0_{-i}\}\cup\{S_{-i}(I_\ell):\ell=1,\ldots,L\}$ is a partition of S_{-i}.

For all $s_{-i}\in S^0_{-i}$, by definition $U_i(s_i,s_{-i})=U_i(t_i,s_{-i})$ for all $t_i\in\operatorname{supp}\sigma_i$. By weak sequential rationality, for all $\ell=1,\ldots,L$, $U_i(s_i,\mu(\cdot|I_\ell)\geq U_i(t_i,\mu(\cdot|I_\ell)$ for all $t_i\in S_i(I_\ell)$; but since $\sigma_i(S_i(I_\ell))=1$, also $U_i(s_i,\mu(\cdot|I_\ell)\geq U_i(\sigma_i,\mu(\cdot|I_\ell)$. Furthermore, fix one such ℓ.

Suppose that $\mu(\{t_{-i}\}|I_\ell)>0$ implies $U_i(s_i,t_{-i})=U_i(\sigma_i,t_{-i})$, and that in addition, for all $t_{-i},t'_{-i}\in\operatorname{supp}\mu(\cdot|I_\ell)$, and all $t_i\in\{s_i\}\cup\operatorname{supp}\sigma_i$, $U_i(t_i,t_{-i})=U_i(t_i,t'_{-i})$. Fix $\bar{t}_{-i}\in\operatorname{supp}\mu(\cdot|I_\ell)$. For all $t_i\in\{s_i\}\cup\operatorname{supp}\sigma_i$,

$$\begin{aligned}U_i\big(t_i,\mu(\cdot|I_\ell)\big)&=\sum_{t_{-i}\in\operatorname{supp}\mu(\cdot|I_\ell)}\mu\big(\{t_{-i}\}|I_\ell\big)U_i(t_i,t_{-i})\\&=\sum_{t_{-i}\in\operatorname{supp}\mu(\cdot|I_\ell)}\mu\big(\{t_{-i}\}|I_\ell\big)U_i(t_i,\bar{t}_{-i})=U_i(t_i,\bar{t}_{-i}).\end{aligned}$$

By weak sequential rationality, $U_i(s_i,\mu(\cdot|I_\ell))\geq U_i(t_i,\mu(\cdot|I_\ell))$ for all $t_i\in S_i(I_\ell)$, so in particular for all $t_i\in\operatorname{supp}\sigma_i$. Thus, $U_i(s_i,\bar{t}_{-i})\geq U_i(t_i,\bar{t}_{-i})$ for all $t_i\in\operatorname{supp}\sigma_i$. By assumption, $U_i(s_i,\bar{t}_{-i})=U_i(\sigma_i,\bar{t}_{-i})$, so it must be that $U_i(s_i,\bar{t}_{-i})=U_i(t_i,\bar{t}_{-i})$, for all $t_i\in\operatorname{supp}\sigma_i$. Since there is $\bar{t}_i\in\operatorname{supp}\sigma_i$ with $\bar{t}_i(I_\ell)\neq s_i(I_\ell)$, we have $\zeta(s_i,\bar{t}_{-i})\neq\zeta(\bar{t}_i,\bar{t}_{-i})$ and $U_i(s_i,\bar{t}_{-i})=U_i(\bar{t}_i,\bar{t}_{-i})$; that is, $(I_\ell,s_i,\bar{t}_i,\bar{t}_{-i})$ is a relevant tie, contradiction.

Suppose instead that $\mu(\{t_{-i}\}|I_\ell)>0$ implies $U_i(s_i,t_{-i})=U_i(\sigma_i,t_{-i})$, but there are $t_{-i},t'_{-i}\in\operatorname{supp}\mu(\cdot|I_\ell)$, and $t_i\in\{s_i\}\cup\operatorname{supp}\sigma_i$ such that $U_i(t_i,t_{-i})\neq U_i(t_i,t'_{-i})$. If $t_i=s_i$, then

$U_i(\sigma_i, t_{-i}) = U_i(s_i, t_{-i}) \neq U_i(s_i, t'_{-i}) = U_i(\sigma_i, t'_{-i})$, so there must be $\bar{t}_i \in \operatorname{supp}\sigma_i$ such that $U_i(\bar{t}_i, t_{-i}) \neq U_i(\bar{t}_i, t'_{-i})$. Thus, it is wlog to take $t_i = \bar{t}_i \in \operatorname{supp}\sigma_i$; thus, $\zeta(t_i, t_{-i}) \neq \zeta(t_i, t'_{-i})$. But then, since $\sigma_i(S_{-i}(I_\ell) \setminus \{s_i\}) = 1$, $(I_\ell, s_i, \sigma_i, t_{-i}, t'_{-i})$ is a nontrivial redundance, contradiction.

To sum up, there exists $t_{-i} \in S_{-i}(I_\ell)$ such that $\mu(\{t_{-i}\}|I_\ell) > 0$ and either $U_i(s_i, t_{-i}) > U_i(\sigma_i, t_{-i})$ or $U_i(s_i, t_{-i}) < U_i(\sigma_i, t_{-i})$. Write $S^+_{-i}(I_\ell)$ and, respectively, $S^-_{-i}(I_\ell)$, for the collection of $t_{-i} \in S_{-i}(I_\ell)$ for which $U_i(s_i, t_{-i}) > U_i(\sigma_i, t_{-i})$ and, respectively, $U_i(s_i, t_{-i}) < U_i(\sigma_i, t_{-i})$. Since, for all $t_{-i} \notin S^+_{-i}(I_\ell) \cup S^-_{-i}(I_\ell)$, either $\mu(\{t_{-i}\}|I_\ell) = 0$ or $U_i(s_i, t_{-i}) = U_i(\sigma_i, t_{-i})$ (or both), $U_i(s_i, \mu(\cdot|I_\ell) \geq U_i(\sigma_i, \mu(\cdot|I_\ell))$ implies that

$$\sum_{t_{-i}\in S^+_{-i}(I_\ell)} \mu(\{t_{-i}\}|I_\ell)\bigl[U_i(s_i, t_{-i}) - U_i(\sigma_i, t_{-i})\bigr] \geq \sum_{t_{-i}\in S^-_{-i}(I_\ell)} \mu(\{t_{-i}\}|I_\ell)\bigl[U_i(\sigma_i, t_{-i}) - U_i(s_i, t_{-i})\bigr] \geq 0, \tag{18}$$

and at least one inequality is strict. Thus, $\sum_{t_{-i}\in S^+_{-i}(I_\ell)} \mu(\{t_{-i}\}|I_\ell)[U_i(s_i, t_{-i}) - U_i(\sigma_i, t_{-i})] > 0$.

Now fix a perturbation $(p^k)_{k\geq 1}$ of μ. For every ℓ, eventually $\sum_{t_{-i}\in S^+_{-i}(I_\ell)} p^k(\{t_{-i}\}|S_{-i}(I_\ell))[U_i(s_i, t_{-i}) - U_i(\sigma_i, t_{-i})] > 0$, so for k large, the quantity

$$\alpha^k_\ell \equiv \frac{\sum_{t_{-i}\in S^-_{-i}(I_\ell)} p^k(\{t_{-i}\}|S_{-i}(I_\ell))\bigl[U_i(\sigma_i, t_{-i}) - U_i(s_i, t_{-i})\bigr]}{\sum_{t_{-i}\in S^+_{-i}(I_\ell)} p^k(\{t_{-i}\}|S_{-i}(I_\ell))\bigl[U_i(s_i, t_{-i}) - U_i(\sigma_i, t_{-i})\bigr]}$$

is well-defined. By equation (18), $\lim_{k\to\infty} \alpha^k_\ell \leq 1$. Let $\beta^k_\ell = \max(\alpha^k(s_{-i}), 1)$, so $\beta^k_\ell \geq 1$ and $\beta^k_\ell \to 1$; let $c = (p^k(S^0_{-i}) + \sum_{m=1}^L [\beta^k_m p^k(S^+_{-i}(I_m)) + p^k(S^-_{-i}(I_m))])^{-1}$. Finally, define $(\tilde{p}^k)_{k\geq 1}$ by

$$\tilde{p}^k(\{t_{-i}\}) = \begin{cases} c \cdot \beta^k_\ell p^k(\{t_{-i}\}) & t_{-i} \in S^+_{-i}(I_\ell) \text{ for some } \ell = 1, \dots, L; \\ c \cdot p^k(\{t_{-i}\}) & \text{otherwise} \end{cases}$$

for every $k \geq 1$ and $t_{-i} \in S_{-i}$. By construction, for every $\ell = 1, \dots, L$ and every $k \geq 1$,

$$\begin{aligned} &\sum_{t_{-i}\in S^+_{-i}(I_\ell)} \tilde{p}^k(\{t_{-i}\}|S_{-i}(I_\ell))\bigl[U_i(s_i, t_{-i}) - U_i(\sigma_i, t_{-i})\bigr] \\ &= \frac{1}{\tilde{p}^k(S_{-i}(I_\ell))} \cdot \sum_{t_{-i}\in S^+_{-i}(I_\ell)} \tilde{p}^k(\{t_{-i}\})\bigl[U_i(s_i, t_{-i}) - U_i(\sigma_i, t_{-i})\bigr] \\ &= \frac{1}{\tilde{p}^k(S_{-i}(I_\ell))} \cdot c \cdot \beta^k_\ell \cdot \sum_{t_{-i}\in S^+_{-i}(I_\ell)} p^k(\{t_{-i}\})\bigl[U_i(s_i, t_{-i}) - U_i(\sigma_i, t_{-i})\bigr] \\ &= \frac{p^k(S_{-i}(I_\ell))}{\tilde{p}^k(S_{-i}(I_\ell))} \cdot c \cdot \beta^k_\ell \cdot \sum_{t_{-i}\in S^+_{-i}(I_\ell)} p^k(\{t_{-i}\}|S_{-i}(I_\ell))\bigl[U_i(s_i, t_{-i}) - U_i(\sigma_i, t_{-i})\bigr] \end{aligned}$$

$$\geq \frac{p^k(S_{-i}(I_\ell))}{\tilde{p}^k(S_{-i}(I_\ell))} \cdot c \cdot \alpha_\ell^k \cdot \sum_{t_{-i} \in S^+_{-i}(I_\ell)} p^k(\{t_{-i}\}|S_{-i}(I_\ell))[U_i(s_i, t_{-i}) - U_i(\sigma_i, t_{-i})]$$

$$= \frac{p^k(S_{-i}(I_\ell))}{\tilde{p}^k(S_{-i}(I_\ell))} \cdot c \cdot \sum_{t_{-i} \in S^-_{-i}(I_\ell)} p^k(\{t_{-i}\}|S_{-i}(I_\ell)\cdot)[U_i(\sigma_i, t_{-i}) - U_i(s_i, t_{-i})]$$

$$= \frac{1}{\tilde{p}^k(S_{-i}(I_\ell))} \cdot c \cdot \sum_{t_{-i} \in S^-_{-i}(I_\ell)} p^k(\{t_{-i}\})[U_i(\sigma_i, t_{-i}) - U_i(s_i, t_{-i})]$$

$$= \frac{1}{\tilde{p}^k(S_{-i}(I_\ell))} \cdot \sum_{t_{-i} \in S^-_{-i}(I_\ell)} \tilde{p}^k(\{t_{-i}\})[U_i(\sigma_i, t_{-i}) - U_i(s_i, t_{-i})]$$

$$= \sum_{t_{-i} \in S^-_{-i}(I_\ell)} \tilde{p}^k(\{t_{-i}\}|S_{-i}(I_\ell))[U_i(\sigma_i, t_{-i}) - U_i(s_i, t_{-i})];$$

hence, $U_i(s_i, \tilde{p}^k(\cdot|S_{-i}(I_\ell)) \geq U_i(\sigma_i, \tilde{p}^k(\cdot|S_{-i}(I_\ell))$. Since this holds for all ℓ, $\{S^0_{-i}\} \cup \{S_{-i}(I_\ell) : \ell = 1, \dots, L\}$ is a partition of S_{-i}, and $U_i(s_i, s_{-i})) = U_i(\sigma_i, s_{-i})$ for all $s_{-i} \in S^0_{-i}$, $U_i(s_i, \tilde{p}^k) \geq U_i(\sigma_i, \tilde{p}^k)$.

It remains to be shown that $\tilde{p}^k$ is a perturbation of μ. Since each $\tilde{p}^k$ has the same support as p^k, $\tilde{p}^k(S_{-i}(I)) > 0$ for all $I \in \mathcal{I}_i$ and $k \geq 1$. Now fix one such I and $s_{-i} \in S_{-i}(I)$ with $\mu(\{s_{-i}\}|I) > 0$. Then eventually $\tilde{p}^k(\{s_{-i}\}) > 0$, and for any other $t_{-i} \in S_{-i}(I)$,

$$\frac{\tilde{p}^k(\{t_{-i}\})}{\tilde{p}^k(\{s_{-i}\})} = \frac{\gamma^k(t_{-i}) \cdot p^k(\{t_{-i}\})}{\gamma^k(s_{-i}) \cdot p^k(\{s_{-i}\})} = \frac{\gamma^k(t_{-i}) \cdot p^k(\{t_{-i}\}|S_{-i}(I))}{\gamma^k(s_{-i}) \cdot p^k(\{s_{-i}\}|S_{-i}(I))} \to \frac{\mu(\{t_{-i}\}|I)}{\mu(\{s_{-i}\}|I)},$$

where $\gamma^k(r_{-i}) = \beta^k_\ell$ if $r_{-i} \in S^+_{-i}(I_\ell)$ for some ℓ, and $\gamma^k(r_{-i}) = 1$ otherwise, so that $\gamma^k(r_{-i}) \to 1$ in either case. This implies that $\tilde{p}^k(\cdot|S_{-i}(I)) \to \mu(\cdot|I)$. *Q.E.D.*

B.4. *Elicitation*

Throughout this section, fix a dynamic game $(N, A, Z, P, (\mathcal{I}_i, u_i)_{i\in N})$, a questionnaire $Q = (I_i, W_i)_{i\in N}$, and an elicitation game $(N \cup \{0\}, A^*, Z^*, P^*, (\mathcal{I}^*_i, u^*_i)_{i\in N\cup\{0\}}, \epsilon)$ for Q.

For $s^* \in S^*$, let $s^*_{-0i} = (s^*_j)_{j\in N\setminus\{i\}}$ and $S^*_{-0i} = \prod_{j\in N\setminus\{i\}} S^*_j$.

LEMMA 1: *$S^*(I^1_i) = S^*$ for every $i \in N$. Furthermore, for all $I_{\bar{s}_i, w_i} \in \mathcal{I}_i$,*

$$S^*(I_{\bar{s}_i,w_i}) = \{i\} \times \{s^*_i : \mathbf{r}_i(s^*_i) = \bar{s}_i, \mathbf{w}_i(s^*_i) = w_i, \mathbf{d}_i(s^*_i) \in S_i(I)\} \times \{s^*_{-0i} : (\mathbf{r}_j(s^*_j))_{j\in N\setminus\{i\}} \in S_{-i}(I)\}. \tag{19}$$

PROOF: $S^*(\phi^*) = S^*(I^1_i) = S^*$ follows immediately from Definition 5. Now consider $I_{\bar{s}_i,w_i} \in \mathcal{I}^*_i$. By definition, $S^*(I_{\bar{s}_i,w_i}) = \bigcup_{h^*\in I_{\bar{s}_i,w_i}} S^*(h^*)$.

CLAIM: *Let $h^* = (n, (\bar{t}_1, v_1), \dots, (\bar{t}_i, v_i), \dots, (\bar{t}_i, v_i), h) \in N \times \prod_{i\in N}(S_i \times W_i) \times (H \setminus Z)$. Then $h^* \in I_{\bar{s}_i,w_i}$ iff $n = i$, $\bar{t}_i = \bar{s}_i$, $v_i = w_i$, $h \in I$, and there is $t_i \in S_i$ such that $(t_i, \bar{t}_{-i}) \in S(h)$.*

PROOF: If $h^* \in I_{\bar{s}_i,w_i}$, then by definition $n = i$, $\bar{t}_i = \bar{s}_i$, $v_i = w_i$, and $h \in I$; moreover, since $h^* \in I_{\bar{s}_i,w_i}$ implies $h^* \in H^*$, the definition of H^* implies that $(n, (\bar{t}_1, v_1), \dots, (\bar{t}_i, v_i), \dots, (\bar{t}_i, v_i), z) \in Z^*$ for some $z \in Z$ such that $h < z$. By the definition of Z^*, $\bar{t}_{-i} \in S_{-i}(z)$, so there is $t_i \in S_i$ is such that $(t_i, \bar{t}_{-i}) \in S(z)$. Since $h < z$, $(t_i, \bar{t}_{-i}) \in S(h)$ as well, as claimed. Conversely, suppose that $n = i$, $\bar{t}_i = \bar{s}_i$, $v_i = w_i$, $h \in I$, and $(t_i, \bar{t}_{-i}) \in S(h)$. Let $z = \zeta(t_i, \bar{t}_{-i})$: then $h < z$ and by construction $\bar{t}_{-i} \in S_{-i}(z)$; hence, $z^* \equiv (i, (\bar{t}_1, v_1), \dots, (\bar{t}_i, v_i), \dots, (\bar{s}_i, w_i), z) \in Z^*$, so $h^* \in H^*$; and since $n = i$, $\bar{t}_i = \bar{s}_i$, $v_i = w_i$, and $h \in I$, $h^* \in I_{\bar{s}_i,w_i}$, as claimed. *Q.E.D.*

Now fix $s^* \in S^*(I_{\bar{s}_i,w_i})$, so $s^* \in S^*(h^*)$ for some $h^* \in I_{\bar{s}_i,w_i}$. By the claim,

$$h^* = \big(i, (\bar{t}_1, v_1), \dots, (\bar{s}_i, w_i), \dots, (\bar{t}_N, v_N), h\big)$$

for some $h \in I$, and there is $t_i \in S_i$ such that $(t_i, \bar{t}_{-i}) \in S(h)$, so $\bar{t}_{-i} \in S_{-i}(h)$. By definition, $s^* \in S^*(h^*)$ then implies that $s_0^*(\phi^*) = i$ and $(\mathbf{r}_j(s_j^*), \mathbf{w}_j(s_j^*)) = s_j^*(I_j^1) = (\bar{t}_j, v_j)$ for $j \in N \setminus \{i\}$, so $(\mathbf{r}_j(s_j^*)_{j\in N\setminus\{i\}} = \bar{t}_{-i} \in S_{-i}(h) \subseteq S_{-i}(I)$. Also, $(\mathbf{r}_i(s_i^*), \mathbf{w}_i(s_i^*)) = s_i^*(I_i^1) = (\bar{s}_i, w_i)$.

In addition, let $h = (a_1, \dots, a_K)$, and consider $k \in \{1, \dots, K\}$ such that $P((a_1, \dots, a_{k-1})) = i$. Let $J \in \mathcal{I}_i$ be such that $(a_1, \dots, a_{k-1}) \in J$, and define $h_{k-1}^* = (i, (\bar{t}_1, v_1), \dots, (\bar{s}_i, w_i), \dots, (\bar{t}_N, v_N), a_1, \dots, a_{k-1})$. As noted above, there is t_i such that $(t_i, \bar{t}_{-i}) \in S(h) \subseteq S((a_1, \dots, a_{k-1}))$. Hence, by the claim, $h_{k-1}^* \in J_{\bar{s}_i,w_i}$. By the definition of $\mathbf{d}_i(\cdot)$, since $s^* \in S^*(h^*)$, $\mathbf{d}_i(s_i^*)(J) = s_i^*(J_{\bar{s}_i,w_i}) = a_k$. By Remark 1, $\mathbf{d}_i(s_i^*) \in S_i(h) \subseteq S_i(I)$. Therefore, s^* belongs to the right-hand side of equation (19).

Conversely, suppose s^* belongs to the right-hand side of equation (19). By assumption $(\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}} \in S_{-i}(I)$ and $\mathbf{d}_i(s_i^*) \in S_i(I)$, so by perfect recall $(\mathbf{d}_i(s_i^*), (\mathbf{r}_j(s_j^*))) \in S(I)$. Hence, there is $h \in I$ such that $(\mathbf{d}_i(s_i^*), (\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}) \in S(h)$. Let

$$h^* \equiv \big(s_0^*(\phi^*), \big(\mathbf{r}_1(s_i^1), \mathbf{w}_1(s_i^*)\big), \dots, \big(\mathbf{r}_i(s_i^*), \mathbf{w}_i(s_i^*)\big), \dots, \big(\mathbf{r}_N(s_N^*), \mathbf{w}_N(s_N^*)\big), h\big).$$

By assumption $s_0^*(\phi^*) = i$, $\mathbf{r}_i(s_i^*) = \bar{s}_i$, and $\mathbf{w}_i(s_i^*) = w_i$. Furthermore, $(\mathbf{d}_i(s_i^*), (\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}) \in S(h)$—that is, one can take $t_i = \mathbf{d}_i(s_i^*)$ in the statement of the claim. Hence, $h^* \in I_{\bar{s}_i,w_i}$. It remains to be shown that $s^* \in S^*(h^*)$.

Write $h^* = (a_1^*, \dots, a_K^*)$, with $K \geq N + 1$. Thus, $h = (a_{N+2}^*, \dots, a_K^*)$.[15] According to the definition, it must be shown that, for all $k = 1, \dots, K$, action a_k^* is specified by s^* at history $(a_1^*, \dots, a_{k-1}^*)$. There are two cases to consider. If $1 \leq k \leq N + 1$, then either $k = 1$, in which case $a_k^* = s_0^*(\phi^*)$ by the definition of h^*, or $(a_1^*, \dots, a_{k-1}^*) = (s_0^*(\phi^*), (\mathbf{r}_1(s_1^*), \mathbf{w}_1(s_1^*)), \dots, (\mathbf{r}_{k-2}(s_{k-2}^*), \mathbf{w}_{k-2}(s_{k-2}^*)) \in I_{k-1}^1$ and, by the definition of h^*, $\mathbf{r}_i(\cdot)$, and $\mathbf{w}_i(\cdot)$, $s_{k-1}^*(I_{k-1}^1) = (\mathbf{r}_{k-1}(s_{k-1}^*), \mathbf{w}_{k-1}(s_{k-1}^*)) = a_k^*$.

If instead $k > N + 1$, then $(a_1^*, \dots, a_{k-1}^*) = (s_0^*(\phi^*), (\mathbf{r}_1(s_1^*), \mathbf{w}_1(s_1^*)), \dots, (\mathbf{r}_N(s_N^*), \mathbf{w}_N(s_N^*)), a_{N+2}^*, \dots, a_{k-1}^*)$, where $h' \equiv (a_{N+2}^*, \dots, a_{k-1}^*) < (a_{N+2}^*, \dots, a_k^*) \leq h$.[16] There are two subcases.

If $P(h') = i$, then also $P^*((a_1^*, \dots, a_{k-1}^*)) = i$, and there exists $J \in \mathcal{I}_i$ such that $h' \in J$. Furthermore, $s_0^*(\phi^*) = i$, $\mathbf{r}_i(s_i^*) = \bar{s}_i$, $\mathbf{w}_i(s_i^*) = w_i$, and $(\mathbf{d}_i(s_i^*), (\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}) \in S(h) \subseteq S(h')$. Therefore, by the claim, $(a_1^*, \dots, a_{k-1}^*) \in J_{\bar{s}_i,w_i}$. Also, by Remark 1, $\mathbf{d}_i(s_i^*) \in S_i(h)$ implies $\mathbf{d}_i(s_i^*)(J) = a_k^*$. We conclude that $s_i^*(J_{\bar{s}_i,w_i}) = \mathbf{d}_i(s_i^*)(J) = a_k^*$.

If instead $P(h') = j \neq i$, then as above there is $J \in \mathcal{I}_j$ with $h' \in J$. In this case, $(\mathbf{d}_i(s_i^*), (\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}) \in S(h) \subseteq S(h')$ implies that $\mathbf{r}_j(s_j^*)(J) = a_k^*$. Moreover, now $P^*((a_1^*,$

[15] $K = N + 1$ corresponds to $h = \phi$.

[16] $k = N + 2$ is also allowed, in which case $(a_{N+2}^*, \dots, a_{k-1}^*) = \phi$.

$\ldots, a^*_{k-1})) = 0$, and $(a^*_1, \ldots, a^*_{k-1})$ is contained in the singleton information set $J^* = \{(a^*_1, \ldots, a^*_{k-1})\} \in \mathcal{I}^*_0$. Now suppose that $a \in A$ is such that

$$\begin{aligned}(a^*_1, \ldots, a^*_{k-1}, a) = \big(&s^*_0(\phi^*), (\mathbf{r}_1(s^1_i), \mathbf{w}_1(s^*_i)), \ldots, (\mathbf{r}_i(s^*_i), \mathbf{w}_i(s^*_i)), \ldots, \\ &(\mathbf{r}_N(s^*_N), \mathbf{w}_N(s^*_N)), a^*_{N+2}, \ldots, a^*_{k-1}, a\big) \\ \in H^*.&\end{aligned}$$

Then $(a^*_1, \ldots, a^*_{k-1}, a) < z^*$ for some $z^* \in Z^*$, and there must exist $z \in Z$ such that

$$z^* = \big(s^*_0(\phi^*), (\mathbf{r}_1(s^1_i), \mathbf{w}_1(s^*_i)), \ldots, (\mathbf{r}_i(s^*_i), \mathbf{w}_i(s^*_i)), \ldots, (\mathbf{r}_N(s^*_N), \mathbf{w}_N(s^*_N)), z\big).$$

This requires that $(h', a) = (a^*_{N+2}, \ldots, a^*_{k-1}, a) < z$. In addition, the definition of Z^* requires that $(\mathbf{r}_j(s^*_j))_{j\in N\setminus\{i\}} \in S_{-i}(z)$ (recall that $s^*_0(\phi^*) = i$), so by Remark 1, in particular $\mathbf{r}_j(s^*_j)(J) = a$. But then $a = a^*_k$. We conclude that $A(J^*) = \{a^*_k\}$, so necessarily $s^*_0(J^*) = a^*_k$, as needed. *Q.E.D.*

For every $s_{-i} \in S_{-i}$, let $[s_{-i}] = \{t^*_{-0i} \in S^*_{-0i} : \forall j \neq i, \mathbf{r}_j(t^*_j) = s_j\}$. The collection $\{[s_{-i}] : s_{-i} \in S_{-i}\}$ partitions S^*_{-0i}. Furthermore, from equation (19),

$$S^*_{-i}(I_{\bar{s}_i,w_i}) = \{i\} \times \bigcup_{s_{-i}\in S_{-i}(I)} [s_{-i}]. \tag{20}$$

For every $i \in N$, $s_i \in S_i$, and $w_i \in W_i$, let $s^*_i(\bar{s}_i, w_i, s_i)$ be the element of S^*_i such that $s^*_i(\bar{s}_i, w_i, s_i)(I^1_i) = (\bar{s}_i, w_i)$ and, for all $I \in \mathcal{I}_i$ and $(\bar{s}'_i, w'_i) \in S_i \times W_i$, $s^*_i(\bar{s}_i, w_i, s_i)(I_{\bar{s}'_i,w'_i}) = s_i(I)$. That is, $s^*_i(\bar{s}_i, w_i, s_i)$ plays $(\bar{s}_i, w_i)$ in the first stage, and then, if called upon to play directly, plays according to s_i at all information sets, including those that follow stage-1 choices different from $(\bar{s}_i, w_i)$.

OBSERVATION 1: $\mathbf{r}_i(s^*_i(\bar{s}_i, w_i, s_i) = \bar{s}_i$, $\mathbf{w}_i(s^*_i(\bar{s}_i, w_i, s_i) = w_i$, and $\mathbf{d}_i(s^*_i(\bar{s}_i, w_i, s_i) = s_i$.

LEMMA 2: *For all $\mu_i \in \Delta(S_{-i}, \mathcal{I}_i)$, there is $\mu^*_i \in \Delta(S^*_{-i}, \mathcal{I}^*_i)$ that agrees with μ_i.*

PROOF: For any $s_{-i} \in S_{-i}$ and $n \in N$, let $s^*_{-i}(n, s_{-i}, w_{-i})$ the element of S^*_{-i} such that $s^*_0 = n$ and, for all $j \notin \{i, 0\}$, $s^*_j(n, s_{-i}, w_{-i}) = s^*_j(s_j, w_j, s_j)$. Let $S^{**}_{-i} = \{s^*_{-i} \in S^*_{-i} : \exists (n, s_{-i}, w_{-i}) \in N \times S_{-i} \times W_{-i} : s^*_{-i} = s^*_{-i}(n, s_{-i}, w_{-i})\}$.

Define $\mu^*_i \in \Delta(S^*_{-i})^{\mathcal{I}^*_i}$ by letting, for every $n \in N$, $w_{-i} \in W_{-i}$, and $s_{-i} \in S_{-i}$,

$$\mu^*_i\big(\{s^*_{-i}(n, s_{-i}, w_{-i})\}|\phi^*\big) = \frac{1}{N \cdot |W_{-i}|}\mu\big(\{s_{-i}\}|\phi\big) \quad \text{and}$$

$$\forall I_{\bar{s}_i,w_i} \in \mathcal{I}^*_i, \quad \mu^*_i\big(\{s^*_{-i}(i, s_{-i}, w_{-i})\}|I_{\bar{s}_i,w_i}\big) = \frac{1}{|W_{-i}|}\mu\big(\{s_{-i}\}|I\big),$$

and then defining $\mu^*_i(E^*|I^*_i) = \sum_{s^*_{-i}\in S^{**}_{-i}\cap E} \mu^*_i(\{s_{-i}\}|I^*_i)$ for all $E \subseteq S^*_{-i}$. The fact that this does in fact define probabilities on S^*_{-i} is immediate; furthermore, $\mu^*_i(S^{**}_{-i}|I^*_i) = 1$ for all $I^*_i \in \mathcal{I}^*_i$.

Let $(p^k)_{k\geq 1}$ be a perturbation of μ_i. Define $(q^k)_{k\geq 1} \subseteq \Delta(S^*_{-i})$ by letting $q^k(\{s^*_{-i}(n, s_{-i}, w_{-i})\}) = \frac{1}{N\cdot|W_{-i}|} p^k(\{s_{-i}\})$ for all $k \geq 1$, $n \in N$, $s_{-i} \in S_{-i}$ and $w_{-i} \in W_{-i}$, and then letting

$q^k(E^*) = \sum_{s^*_{-i} \in S^{**}_{-i} \cap E} q^k(\{s_{-i}\})$ for all $E \subseteq S^*_{-i}$. Again, this does in fact define probabilities on S^*_{-i}, and $q^k(S^{**}_{-i}) = 1$.

Then $q^k(\{s^*_{-i}(n, s_{-i}, w_{-i}\}) = \frac{1}{N\cdot|W_{-i}|} p^k(\{s_{-i}\}) \to \frac{1}{N\cdot|W_{-i}|}\mu(\{s_{-i}\}|\phi) = \mu^*_i(\{s^*_{-i}(n, s_{-i}, w_{-i})\}|\phi^*)$. Furthermore, for all $I \in \mathcal{I}_i$ and $(\bar{s}_i, w_i) \in S_i \times W_i$, for all $s_{-i} \in S_{-i}(I)$ and $w_{-i} \in W_{-i}$, by equation (19) and the fact that $q^k(S^{**}_{-i}) = 1$,

$$\begin{aligned}
&q^k\big(\{s^*_{-i}(i, s_{-i}, w_{-i})\}|S^*_{-i}(I_{\bar{s}_i,w_i})\big)\\
&\quad= \frac{q^k\big(\{s^*_{-i}(i, s_{-i}, w_{-i})\}\big)}{\sum\limits_{t_{-i}\in S_{-i}(I)} q^k\big(\{i\}\times[t_{-i}]\big)} = \frac{q^k\big(\{s^*_{-i}(i, s_{-i}, w_{-i})\}\big)}{\sum\limits_{t_{-i}\in S_{-i}(I),\tilde{w}_{-i}\in W_{-i}} q^k\big(s^*_{-i}(i, t_{-i}, w_{-i})\big)}\\
&\quad= \frac{\frac{1}{N\cdot|W_{-i}|}p^k\big(\{s_{-i}\}\big)}{\sum\limits_{t_{-i}\in S_{-i}(I),\tilde{w}_{-i}\in W_{-i}} \frac{1}{N\cdot|W_{-i}|}p^k\big(\{t_{-i}\}\big)} = \frac{1}{|W_{-i}|}p^k\big(\{s_{-i}\}|S_{-i}(I)\big)\\
&\quad\to \frac{1}{|W_{-i}|}\mu_i\big(\{s_{-i}\}|I\big) = \mu^*_i\big(\{s^*_{-i}(n, s_{-i}, w_{-i})\}|I_{\bar{s}_i,w_I}\big).
\end{aligned}$$

Thus, μ^*_i is a CCPS. Finally, I show that μ^*_i agrees with μ_i. Fix $s_{-i} \in S_{-i}$; for $I^*_i = \phi^*$,

$$\begin{aligned}
&\mu^*_i\big(\{t^*_{-i} : t^*_0 = n, \bar{s}_j(t^*_j) = s_j\ \forall j \in N\setminus\{i\}\}|\phi^*\big)\\
&\quad= \sum_{w_{-i}\in W_{-i}} \mu^*_i\big(\{s^*_{-i}(n, s_{-i}, w_{-i})\}|\phi^*\big)\\
&\quad= \sum_{w_{-i}\in W_{-i}} \frac{1}{N\cdot|W_{-i}|}\mu_i\big(\{s_{-i}\}|\phi\big) = \frac{1}{N}\mu_i\big(\{s_{-i}\}|\phi\big),
\end{aligned}$$

where the first equality follows from $\mu^*_i(S^{**}_{-i}|\phi^*) = 1$. For $I^*_i = I_{\bar{s}_i,w_i} \in \mathcal{I}^*_i$,

$$\begin{aligned}
&\mu^*_i\big(\{t^*_{-i} : t^*_0 = i, \bar{s}_j(t^*_j) = s_j\ \forall j \in N\setminus\{i\}\}|I_{\bar{s}_i,w_i}\big)\\
&\quad= \sum_{w_{-i}\in W_{-i}} \mu^*_i\big(\{s^*_{-i}(i, s_{-i}, w_{-i})\}|I_{\bar{s}_i,w_i}\big)\\
&\quad= \sum_{w_{-i}\in W_{-i}} \frac{1}{|W_{-i}|}\mu_i\big(\{s_{-i}\}|I\big) = \mu_i\big(\{s_{-i}\}|I\big),
\end{aligned}$$

which completes the proof. *Q.E.D.*

LEMMA 3: *Consider a CCPS $\mu^*_i \in \Delta(S^*_{-i}, \mathcal{I}^*_i)$ that agrees with μ_i. Then*:

(i) *For every perturbation $(q^k)_{k\geq 1}$ of μ^*_i, there exists a finite index $\kappa \geq 1$ such that $q^\ell(\{i\}\times S^*_{-0i}) > 0$ for all $\ell \geq \kappa$, and the sequence $(p^k)_{k\geq 1} \in \Delta(S_{-i})^{\mathbb{N}}$ defined by*

$$p^k\big(\{s_{-i}\}\big) = q^{k+\kappa-1}\big(\{i\}\times[s_{-i}] \mid \{i\}\times S^*_{-0i}\big), \quad s_{-i} \in S_{-i}, k \geq 1 \tag{21}$$

is a perturbation of μ_i.

(ii) *For every perturbation* $(p^k)_{k\geq 1}$ *of* μ_i, *there is a perturbation* $(q^k)_{k\geq 1}$ *of* μ_i^* *that satisfies equation* (21) *with* $\kappa = 1$.

PROOF: For (i), by equation (10) and the fact that $(q^k)_{k\geq 1}$ is a perturbation of μ_i^*, $\mu_i^*(\{i\} \times S^*_{-0i}|\phi^*) = \frac{1}{N} = \lim_k q^k(\{i\} \times S^*_{-0i})$; this implies that there is $\kappa \geq 1$ such that $q^k(\{i\} \times S^*_{-0i}) > 0$ for all $k \geq \kappa$. Henceforth, to reduce notational clutter, I assume that in fact $\kappa = 1$; the argument goes through unmodified if $\kappa > 1$, simply replacing q^k with $q^{k+\kappa-1}$.

Fix $I \in \mathcal{I}_i$. Then, for every $k \geq 1$, fixing an arbitrary $(\bar{s}_i, w_i) \in A(I_i^1) = S_i \times W_i$,

$$\begin{aligned} p^k\big(S_{-i}(I)\big) &= \sum_{s_{-i}\in S_{-i}(I)} q^k\big(\{i\} \times [s_{-i}]|\{i\} \times S^*_{-i0}\big) \\ &= q^k\big(S^*_{-i}(I_{\bar{s}_i,w_i})|\{i\} \times S^*_{-0i}\big) \geq q^k\big(S^*_{-i}(I_{\bar{s}_i,w_i})\big) > 0; \end{aligned}$$

the last equality follows from equation (20), and the strict inequality from the assumption that $(q^k)_{k\geq 1}$ is a perturbation of μ_i^*. Also, for every $s_{-i} \in S_{-i}(I)$, since by equation (20) $\{i\} \times [s_{-i}] \subseteq S^*_{-i}(I_{\bar{s}_i,w_i})$,

$$\begin{aligned} \lim_{k\to\infty} \frac{p^k\big(\{s_{-i}\}\big)}{p^k\big(S_{-i}(I)\big)} &= \lim_{k\to\infty} \frac{q^k\big(\{i\} \times [s_{-i}]|\{i\} \times S^*_{-0i}\big)}{q^k\big(S^*_{-i}(I_{\bar{s}_i,w_i})|\{i\} \times S^*_{-0i}\big)} = \lim_{k\to\infty} \frac{q^k\big(\{i\} \times [s_{-i}]\big)}{q^k\big(S^*_{-i}(I_{\bar{s}_i,w_i})\big)} \\ &= \mu_i^*\big(\{i\} \times [s_{-i}]|I_{\bar{s}_i,w_i}\big) = \mu_i\big(\{s_{-i}\}|I\big): \end{aligned}$$

the third equality follows from the assumption that $(q^k)_{k\geq 1}$ is a perturbation of μ_i^*, and the last from agreement, that is, equation (11) in Definition 6.

As for prior beliefs, for every $s_{-i} \in S_{-i}$,

$$\begin{aligned} \lim_{k\to\infty} p^k\big(\{s_{-i}\}\big) &= \lim_{k\to\infty} q^k\big(\{i\} \times [s_{-i}]|\{i\} \times S^*_{-0i}\big) \\ &= \lim_{k\to\infty} \frac{q^k\big(\{i\} \times [s_{-i}]\big)}{q^k\big(\{i\} \times S^*_{-0i}\big)} = \frac{\lim_{k\to\infty} q^k\big(\{i\} \times [s_{-i}]\big)}{\lim_{k\to\infty} q^k\big(\{i\} \times S^*_{-0i}\big)} \\ &= \frac{\mu_i^*\big(\{i\} \times [s_{-i}]|\phi^*\big)}{\mu_i^*\big(\{i\} \times S^*_{-0i}|\phi^*\big)} = \frac{\frac{1}{N}\mu_i\big(\{s_{-i}\}|\phi\big)}{\frac{1}{N}} = \mu_i\big(\{s_{-i}\}|I\big): \end{aligned}$$

the third equality holds because $\lim_{k\to\infty} q^k(\{i\} \times S^*_{-0i}) = \mu_i^*(\{i\} \times S^*_{-0i}|\phi^*) > 0$; the fourth follows from the definition of perturbation, and the fifth from equation (10).

For (ii), for every $I^* \in \mathcal{I}_i^* \cup \{\phi^*\}$, let

$$\rho\big(s^*_{-i}; I^*\big) = \begin{cases} \dfrac{\mu_i^*\big(\{s^*_{-i}\}|I^*\big)}{\mu_i^*\big(\{s_0^*\} \times [s_{-i}]|I^*\big)} & s^*_{-0i} \in [s_{-i}], \mu_i^*\big(\{s_0^*\} \times [s_{-i}]|I^*\big) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Fix $s^*_{-i} \in S^*_{-i}$ and let $s_{-i} = (\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}$; thus, $s^*_{-0i} \in [s_{-i}]$. Suppose $\mu_i^*(\{s_0^*\} \times [s_{-i}]|I^*) > 0$ and $\mu_i^*(\{s_0^*\} \times [s_{-i}]|J^*) > 0$ for distinct $I^*, J^* \in \mathcal{I}_i^*$. Since $\mu_i^*(S^*_{-i}(I^*)|I^*) = \mu_i^*(S^*_{-i}(J^*)|J^*) = 1$, $\{s_0^*\} \times [s_{-i}] \cap S^*_{-i}(I^*) \neq \emptyset$ and $\{s_0^*\} \times [s_{-i}] \cap S^*_{-i}(J^*) \neq \emptyset$, so by equation (19), $s^*_{-i} \in \{s_0^*\} \times$

$[s_{-i}] \subseteq S^*_{-i}(I^*) \cap S^*_{-i}(J^*)$.[17] Finally, fix a perturbation $(r^k)_{k\geq 1}$ of μ^*_i. Then $r^k(S^*_{-i}(I^*)) > 0$ for all k, and $r^k(\{s^*_0\} \times [s_{-i}] | S^*_{-i}(I^*)) \to \mu^*_i(\{s^*_0\} \times [s_{-i}] | I^*) > 0$, so

$$\rho(s^*_{-i}; I^*) = \frac{\lim_{k\to\infty} r^k(\{s^*_{-i}\} | S^*_{-i}(I^*))}{\lim_{k\to\infty} r^k(\{s^*_0\} \times [s_{-i}] | S^*_{-i}(I^*))} = \lim_{k\to\infty} \frac{r^k(\{s^*_{-i}\} | S^*_{-i}(I^*))}{r^k(\{s^*_0\} \times [s_{-i}] | S^*_{-i}(I^*))}$$
$$= \lim_{k\to\infty} r^k(\{s^*_{-i}\} | \{s^*_0\} \times [s_{-i}]).$$

By a similar argument, $\rho(s^*_{-i}; J^*) = \lim_{k\to\infty} r^k(\{s^*_{-i}\} | \{s^*_0\} \times [s_{-i}])$. Therefore, $\rho(s^*_{-i}; I^*) = \rho(s^*_{-i}; J^*)$.

Now define $(q^k)_{k\geq 1} \in \Delta(S^*_{-i})^{\mathbb{N}}$ as follows: for every $s^*_{-i} \in S^*_{-i}$, again let $s_{-i} = (\mathbf{r}_j(s^*_j))_{j\in N\setminus\{i\}}$ and

$$q^k(\{s^*_{-i}\}) = \begin{cases} p^k(\{s_{-i}\}) \cdot \dfrac{1}{N} \cdot \rho(s^*_{-i}; I^*) & \mu^*_i(\{s^*_0\} \times [s_{-i}] | I^*) > 0 \text{ for some } I^* \in \mathcal{I}_i; \\ \dfrac{p^k(\{s_{-i}\})}{N \cdot |[s_{-i}]|} & \text{otherwise.} \end{cases}$$

By the preceding argument, this definition is well posed. Furthermore, fix $j \in N$ and $s_{-i} \in S_{-i}$. Suppose first that $\mu^*_i(\{j\} \times [s_{-i}] | I^*) > 0$ for some $I^* \in \mathcal{I}^*_i$. Then

$$\sum_{s^*_{-i}\in\{j\}\times[s_{-i}]} q^k(\{s^*_{-i}\}) = \sum_{s^*_{-i}\in\{j\}\times[s_{-i}]} p^k(\{s_{-i}\}) \cdot \frac{1}{N} \cdot \rho(s^*_{-i}; I^*)$$
$$= \sum_{s^*_{-i}\in\{j\}\times[s_{-i}]} p^k(\{s_{-i}\}) \cdot \frac{1}{N} \cdot \frac{\mu^*_i(\{s^*_{-i}\} | I^*)}{\mu^*_i(\{s^*_0\} \times [s_{-i}] | I^*)}$$
$$= \frac{p^k(\{s_{-i}\})}{\mu^*_i(\{j\} \times [s_{-i}] | I^*)} \cdot \frac{1}{N} \cdot \sum_{s^*_{-i}\in\{j\}\times[s_{-i}]} \mu^*_i(\{s^*_{-i}\} | I^*) = \frac{1}{N} p^k(\{s_{-i}\}).$$

If instead $\mu^*_i([s_{-i}] | I^*) = 0$ for all I^*, then

$$\sum_{s^*_{-i}\in\{j\}\times[s_{-i}]} q^k(\{s^*_{-i}\}) = \sum_{s^*_{-i}\in\{j\}\times[s_{-i}]} p^k(\{s_{-i}\}) \cdot \frac{1}{N \cdot |[s_{-i}]|} = \frac{1}{N} p^k(\{s_{-i}\}).$$

Therefore, for all $j \in N$ and s_{-i}, $q^k(\{j\} \times [s_{-i}]) = \frac{1}{N} p^k(\{s_{-i}\})$. This implies that $q^k(S^*_{-i}) = 1$, so $q^k \in \Delta(S^*_{-i})$, and furthermore,

$$q^k(\{i\} \times [s_{-i}] | \{i\} \times S^*_{-0i}) = \frac{q^k(\{i\} \times [s_{-i}])}{\sum_{t_{-i}\in S_i} q^k(\{i\} \times [t_{-i}])} = \frac{\frac{1}{N} p^k(\{s_{-i}\})}{\sum_{t_{-i}\in S_i} \frac{1}{N} p^k(\{t_{-i}\})} = p^k(\{s_{-i}\}),$$

that is, equation (21) holds.

[17]This implies that if, for example, $I^* = I_{s_i,w_i}$ for some $(s_i, w_i) \in S_i \times W_i$, then necessarily $s^*_0 = i$; if instead $I^* \in \{\phi^*, I^1_i\}$, this need not hold and is similar for J^*. However, this difference is immaterial to the argument in this paragraph.

It remains to be shown that $(q^k)_{k\geq 1}$ is a perturbation of μ_i^*. For every $I^* \in \mathcal{I}_i^*$, either $I^* \in \{\phi^*, I_i^1\}$, in which case trivially $q^k(S_{-i}^*(I^*)) = q^k(S_{-i}^*) = 1$, or $I^* = I_{\bar{s}_i, w_i}$ for some $(\bar{s}_i, w_i) \in S_i \times W_i$ and $I \in \mathcal{I}_i$. Since $(p^k)_{k\geq 1}$ is a perturbation of μ_i, $p^k(S_{-i}(I)) > 0$ for all k. For each $k \geq 1$, there must be $s_{-i} \in S_{-i}(I)$, possibly depending on k, with $p^k(\{s_{-i}\}) > 0$. Since $q^k(\{i\} \times [s_{-i}] | \{i\} \times S_{-0i}^*) = p^k(\{s_{-i}\}) > 0$, also $q^k(\{i\} \times [s_{-i}]) > 0$. Thus, by equation (20), $q^k(S_{-i}^*(I_{\bar{s}_i, w_i})) \geq q^k(\{i\} \times [s_{-i}]) > 0$.

Now consider $I^* \in \{\phi^*, I_i^1\}$. Fix $s_{-i}^* \in S_{-i}^*$ and let $s_{-i} = (\mathbf{r}_j(s_j^*))_{j \in N\setminus\{i\}}$. If $\mu_i^*(\{s_0^*\} \times [s_{-i}] | I^*) > 0$,

$$
\begin{aligned}
q^k(\{s_{-i}^*\}) &= p^k(\{s_{-i}\}) \cdot \frac{1}{N} \cdot \frac{\mu_i^*(\{s_{-i}^*\} | I^*)}{\mu_i^*(\{s_0^*\} \times [s_{-i}] | I^*)} \\
&\to \mu_i(\{s_{-i}\} | \phi) \cdot \frac{1}{N} \cdot \frac{\mu_i^*(\{s_{-i}^*\} | I^*)}{\mu_i^*(\{s_0^*\} \times [s_{-i}] | I^*)} \\
&= \mu_i(\{s_{-i}\} | \phi) \cdot \frac{1}{N} \cdot \frac{\mu_i^*(\{s_{-i}^*\} | I^*)}{\frac{1}{N} \mu_i(\{s_{-i}\} | \phi)} = \mu_i^*(\{s_{-i}^*\} | I^*);
\end{aligned}
$$

the second equality follows from the fact that μ_i^* agrees with μ_i. If instead $\mu_i^*(\{s_0^*\} \times [s_{-i}] | I^*) = 0$, then a fortiori $\mu_i^*(\{s_{-i}^*\} | I^*) = 0$, and by agreement with μ_i also $\mu_i(\{s_{-i}\} | \phi) = 0$, so

$$
q^k(\{s_{-i}^*\}) = p^k(\{s_{-i}\}) \cdot \frac{1}{N} \cdot c \to \mu_i(\{s_{-i]}\} | \phi) \cdot \frac{1}{N} \cdot c = 0 = \mu_i^*(\{s_{-i}^*\} | I^*);
$$

here, $c = \rho(s_{-i}^*; J^*)$ if there exists $J^* \in \mathcal{I}_i^*$ with $\mu_i^*(\{s_0^*\} \times [s_{-i}] | J^*) > 0$, and $c = \frac{1}{\|[s_{-i}]\|}$ otherwise, but since c is independent of k, its value is immaterial to the argument.

Finally, suppose $I^* = I_{\bar{s}_i, w_i}$ for some $I \in \mathcal{I}_i$ and $(\bar{s}_i, w_i) \in S_i \times W_i$. Fix $s_{-i}^*, t_{-i}^* \in S_{-i}^*(I^*)$, with $\mu_i^*(\{t_{-i}^*\} | I^*) > 0$. By the definition of the elicitation game, $s_0^* = t_0^* = i$. Let $s_{-i} = (\mathbf{r}_j(s_j^*))_{j \in N\setminus\{i\}}$ and $t_{-i} = (\mathbf{r}_j(t_j^*))_{j \in N\setminus\{i\}}$. Thus, $\mu_i^*(\{i\} \times [t_{-i}] | I^*) > 0$, and since μ_i^* agrees with μ_i, $\mu(\{t_{-i}\} | I) > 0$. Then, for all k large, $p^k(\{t_{-i}\}) > 0$. Moreover, $\rho(t_{-i}^*; I^*) = \frac{\mu_i^*(\{t_{-i}^*\} | I^*)}{\mu_i^*(\{i\} \times [t_{-i}] | I^*)} > 0$, and so, for k large, $q^k(\{t_{-i}^*\}) = p^k(\{t_{-i}\}) \cdot \frac{1}{N} \cdot \rho(t_{-i}^*; I^*) > 0$ as well.

First, suppose $\mu_i^*(\{i\} \times [s_{-i}] | I^*) > 0$, so, since μ_i^* agrees with μ_i, $\mu_i(\{s_{-i}\} | I) > 0$. Then

$$
\begin{aligned}
\frac{q^k(\{s_{-i}^*\})}{q^k(\{t_{-i}^*\})} &= \frac{p^k(\{s_{-i}\}) \cdot \frac{1}{N} \cdot \rho(s_{-i}^*; I^*)}{p^k(\{t_{-i}\}) \cdot \frac{1}{N} \cdot \rho(t_{-i}^*; I^*)} = \frac{p^k(\{s_{-i}\}) \cdot \dfrac{\mu_i^*(\{s_{-i}^*\} | I^*)}{\mu_i^*([s_{-i}] | I^*)}}{p^k(\{t_{-i}\}) \cdot \dfrac{\mu_i^*(\{t_{-i}^*\} | I^*)}{\mu_i^*([t_{-i}] | I^*)}} \\
&= \frac{p^k(\{s_{-i}\} | S_{-i}(I)) \cdot \dfrac{\mu_i^*(\{s_{-i}^*\} | I^*)}{\mu_i(\{s_{-i}\} | I)}}{p^k(\{t_{-i}\} | S_{-i}(I)) \cdot \dfrac{\mu_i^*(\{t_{-i}^*\} | I^*)}{\mu_i(\{t_{-i}\} | I)}} \to \frac{\mu_i^*(\{s_{-i}^*\} | I^*)}{\mu_i^*(\{t_{-i}^*\} | I^*)} :
\end{aligned}
$$

the last equality follows because μ_i^* agrees with μ_i, and the limit statement from the assumption that $(p^k)_{k\geq 1}$ is a perturbation of μ_i.

If instead $\mu_i^*(\{i\}\times[s_{-i}]|I^*)=0$, then by agreement $\mu_i(\{s_{-i}\}|I)=0$ as well, so

$$\frac{q^k(\{s^*_{-i}\})}{q^k(\{t^*_{-i}\})} \le \frac{q^k(\{i\}\times[s_{-i}])}{q^k(\{t^*_{-i}\})} \le \frac{q^k(\{i\}\times[s_{-i}]|\{i\}\times S^*_{-0i})}{q^k(\{t^*_{-i}\})} = \frac{p^k(\{s_{-i}\})}{p^k(\{t_{-i}\})\cdot\frac{1}{N}\cdot\rho(t^*_{-i};I^*)}$$

$$= \frac{N}{\rho(t^*_{-i};I^*)}\cdot\frac{p^k(\{s_{-i}\}|S_{-i}(I))}{p^k(\{t_{-i}\}|S_{-i}(I))} \to \frac{N}{\rho(t^*_{-i};I^*)}\cdot\frac{\mu_i(\{s_{-i}\}|I)}{\mu_i(\{t_{-i}\}|I)} = 0 = \frac{\mu_i^*(\{s^*_{-i}\}|I^*)}{\mu_i^*(\{t^*_{-i}\}|I^*)}.$$

The first equality is from equation (21); the limit statement follows because $(p^k)_{k\ge1}$ is a perturbation of μ_i, and the last equality follows from $\mu_i^*(\{s^*_{-i}\}|I^*)\le\mu_i^*(\{i\}\times[s_{-i}]|I^*)=0$.

To sum up, in each case $\frac{q^k(\{s^*_{-i}\})}{q^k(\{t^*_{-i}\})}\to\frac{\mu_i^*(\{s^*_{-i}\}|I^*)}{\mu_i^*(\{t^*_{-i}\}|I^*)}$ for every $s^*_{-i}\in S^*_{-i}(I^*)$. Therefore,

$$q^k(\{s^*_{-i}\}|S^*_{-i}(I^*)) = \frac{q^k(\{s^*_{-i}\})}{\sum_{r^*_{-i}\in S^*_{-i}(I^*)} q^k(\{r^*_{-i}\})} = \frac{\dfrac{q^k(\{s^*_{-i}\})}{q^k(\{t^*_{-i}\})}}{\sum_{r^*_{-i}\in S^*_{-i}(I^*)}\dfrac{q^k(\{r^*_{-i}\})}{q^k(\{t^*_{-i}\})}}$$

$$\to \frac{\dfrac{\mu_i^*(\{s^*_{-i}\}|I^*)}{\mu_i^*(\{t^*_{-i}\}|I^*)}}{\sum_{r^*_{-i}\in S^*_{-i}(I^*)}\dfrac{\mu_i^*(\{r^*_{-i}\}|I^*)}{\mu_i^*(\{t^*_{-i}\}|I^*)}} = \frac{\mu_i^*(\{s^*_{-i}\}|I^*)}{\sum_{r^*_{-i}\in S^*_{-i}(I^*)}\mu_i^*(\{r^*_{-i}\}|I^*)} = \mu_i^*(\{s^*_{-i}\}|I^*),$$

where the second equality follows from dividing numerator and denominator by $q^k(\{t^*_{-i}\})>0$, and the third by multiplying both by $\mu_i^*(\{t^*_{-i}\}|I^*)>0$. *Q.E.D.*

Now rewrite the strategic-form payoff function in the elicitation game as follows. Fix $s^*\in S^*$, and let $z^*=\zeta^*(s^*)$. By the definition of the maps $\mathbf{r}_j(\cdot)$ and $\mathbf{w}_j(\cdot)$ for all $j\in N$, letting $n=s_0^*(\phi^*)$, $z^*=(n,(\mathbf{r}_j(s_j^*),\mathbf{w}_j(s_j^*))_{j\in N},z)\in Z^*$, where $\mathbf{r}_j(s_j^*)\in S_j(z)$ for all $j\in N\setminus\{n\}$. In addition, write $z=(a_1,\dots,a_L)$, fix $K\in\{1,\dots,L-1\}$, and let $h=(a_1,\dots,a_K)$. Suppose that $P(h)=n$, so $h\in I\in\mathcal{I}_n$. Then $h^*\equiv(n,(\mathbf{r}_j(s_j^*),\mathbf{w}_j(s_j^*))_{j\in N},h)\in H^*$, $P^*(h^*)=n$, and $h^*\in I_{\bar{s}_n,w_n}$; then since $s^*\in S^*(z^*)$, $s_n^*(I_{\bar{s}_n,w_n})=a_{K+1}$. But by equation (9), $\mathbf{d}_n(s_n^*)(I)=s_n^*(I_{\bar{s}_n,w_n})=a_{K+1}$. Thus, for all K such that $P((a_1,\dots,a_K))=n$, if $(a_1,\dots,a_K)\in I\in\mathcal{I}_n$ then $\mathbf{d}_n(s_n^*)(I)=a_{K+1}$. By Remark 1, $\mathbf{d}_n(s_n^*)\in S_n(z)$, and so $(\mathbf{d}_n(s_n^*),(\mathbf{r}_j(s_j^*))_{j\in N\setminus\{n\}})\in S(z)$, that is, $z=\zeta(\mathbf{d}_n(s_n^*),(\mathbf{r}_j(s_j^*))_{j\in N\setminus\{n\}})$. With this, for every $i\in N$, if $n\ne i$ then $U_i^*(s^*)=0$; if $n=i$, since $u_i(z)=U_i(\mathbf{d}_i(s_i^*),(\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}})$,

$$\begin{aligned}U_i^*(s^*) &= u_i^*(z^*)\\ &= \frac{1}{3}U_i(\mathbf{d}_i(s_i^*),(\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}) + \frac{1}{3}B_i(\mathbf{w}_i(s_i^*),(\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}})\\ &\quad + \frac{1}{3}\epsilon\cdot 1_{\mathbf{r}_i(s_i^*)\in S_i(\zeta(\mathbf{d}_i(s_i^*),(\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}))}.\end{aligned}$$

This emphasizes that, if i is selected, her payoff depends on s^*_{-0i} only through $(\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}$, the profile of coplayers' reported strategies. Now $\mathbf{r}_i(s_i^*)\in S_i(\zeta(\mathbf{d}_i(s_i^*),(\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}))$ if and only if $\zeta(\mathbf{r}_i(s_i^*), (\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}}) = \zeta(\mathbf{d}_i(s_i^*), (\mathbf{r}_j(s_j^*))_{j\in N\setminus\{i\}})$. Therefore, for all $s_i^* \in S_i^*$ and $q \in \Delta(S_{-i}^*)$,

$$U_i^*(s_i^*, q) = \frac{1}{3}\sum_{s_{-i}\in S_{-i}} q(\{i\}\times[s_{-i}])U_i(\mathbf{d}_i(s_i^*), s_{-i}) + \frac{1}{3}\sum_{s_{-i}\in S_{-i}} q(\{i\}\times[s_{-i}])\cdot B_i(\mathbf{w}_i(s_i^*), s_{-i}) + \frac{1}{3}\epsilon\cdot \sum_{\substack{s_{-i}\in S_{-i}:\\ \zeta(\mathbf{d}_i(s_i^*),s_{-i})=\zeta(\mathbf{r}_i(s_i^*),s_{-i})}} q(\{i\}\times[s_{-i}]). \tag{22}$$

PROOF OF THEOREM 2: Throughout, adopt the notation and definitions in the statement. The existence of a CCPS that agrees with μ_i is established in Lemma 2. Now assume that s_i^* is structurally rational given a CCPS μ_i^* that agrees with μ_i.

Let $s_i = \mathbf{d}_i(s_i^*)$, $\bar{s}_i = \mathbf{r}_i(s_i^*)$, and $w_i = \mathbf{w}_i(s_i^*)$. Also let $\hat{s}_i^* = s_i^*(s_i, w_i, s_i)$.

CLAIM: *For every $z \in Z$, $s_i \in S_i(z)$ implies $\bar{s}_i \in S_i(z)$—that is, s_i and $\bar{s}_i$ are* realization-equivalent.

PROOF: Suppose that, for some $z \in Z$, $s_i \in S_i(z)$ but $\bar{s}_i \notin S_i(z)$. Then $\mathbf{d}_i(s_i^*) = \mathbf{d}_i(\hat{s}_i^*) = \mathbf{r}_i(\hat{s}_i^*)$ and $\mathbf{w}_i(s_i^*) = \mathbf{w}_i(\hat{s}_i^*)$, so for all $q \in \Delta(S_{-i}^*)$ the first and second terms in equation (22) for $U_i^*(s_i^*, q)$ and $U_i^*(\hat{s}_i^*, q)$ are the same. Hence,

$$\begin{aligned} &U_i^*(s_i^*, q) - U_i^*(\hat{s}_i^*, q) \\ &= \frac{1}{3}\epsilon\Bigg(\sum_{\substack{s_{-i}\\ \zeta(\mathbf{d}_i(s_i^*),s_{-i})=\zeta(\mathbf{r}_i(s_i^*),s_{-i})}} q(\{i\}\times[s_{-i}]) - \sum_{\substack{s_{-i}\\ \zeta(\mathbf{d}_i(\hat{s}_i^*),s_{-i})=\zeta(\mathbf{r}_i(\hat{s}_i^*),s_{-i})}} q(\{i\}\times[s_{-i}])\Bigg) \\ &= \frac{1}{3}\epsilon\Bigg(\sum_{\substack{s_{-i}:\\ \zeta(s_i,s_{-i})=\zeta(\bar{s}_i,s_{-i})}} q(\{i\}\times[s_{-i}]) - \sum_{\substack{s_{-i}\\ \zeta(s_i,s_{-i})=\zeta(s_i,s_{-i})}} q(\{i\}\times[s_{-i}])\Bigg) \\ &= \frac{1}{3}\epsilon\Bigg(\sum_{\substack{s_{-i}:\\ \zeta(s_i,s_{-i})=\zeta(\bar{s}_i,s_{-i})}} q(\{i\}\times[s_{-i}]) - 1\Bigg). \end{aligned}$$

Fix an arbitrary $t_{-i} \in S_{-i}$ such that $(s_i, t_{-i}) \in S(z)$. It must be the case that $(\bar{s}_i, t_{-i}) \notin S(z)$, for otherwise $\bar{s}_i \in S_i(z)$, a contradiction. Let h be the last common prefix of z and $\zeta(\bar{s}_i, t_{-i})$, that is, the longest nonterminal history such that $h < z$ and $h < \zeta(\bar{s}, t_{-i})$. Then $P(h) = i$; let $h \in I \in \mathcal{I}_i$. Then $s_i, \bar{s}_i \in S_i(I)$ and $s_i(I) \neq \bar{s}_i(I)$. Hence, for all $s_{-i} \in S_{-i}(I)$, $\zeta(s_i, s_{-i}) \neq \zeta(\bar{s}_i, s_{-i})$. It follows that

$$\begin{aligned} U_i^*(s_i^*, q) - U_i^*(\hat{s}_i^*, q) &= \frac{1}{3}\epsilon\Bigg(\sum_{\substack{s_{-i}\\ \zeta(s_i,s_{-i})=\zeta(\bar{s}_i,s_{-i})}} q(\{i\}\times[s_{-i}]) - 1\Bigg) \\ &\leq -\frac{1}{3}\epsilon\sum_{s_{-i}\in S_{-i}(I)} q(\{i\}\times[s_{-i}]). \end{aligned}$$

Finally, for any perturbation $(q^k)_{k\geq 1}$ of μ_i^*, by equation (19),

$$0 < q^k(S_{-i}^*(I_{\bar{s}_i,w_i}) = q^k\Bigg(\{i\} \times \bigcup_{s_{-i}\in S_{-i}(I)} [s_{-i}]\Bigg) = \sum_{s_{-i}\in S_{-i}(I)} q^k\big(\{i\}\times[s_{-i}]\big).$$

Therefore, for all perturbations $\{q^k\}_{k\geq 1}$ of μ_i^*, and all k, $U_i^*(\hat{s}_i^*, q^k) > U_i^*(s_i^*, q^k)$. But then s_i^* is not structurally rational for μ_i^*, a contradiction. Thus, $\bar{s}_i \in S_i(z)$ as well. *Q.E.D.*

Now consider part (iii) of the theorem. Fix $z \in Z$. By the claim, if $s_i \in S_i(z)$, then $\bar{s}_i \in S_i(z)$ as well. Conversely, suppose that $\bar{s}_i \in S_i(z)$. Let $s_{-i} \in S_{-i}(z)$, so $(\bar{s}_i, s_{-i}) \in S(z)$ by Remark 1. Thus, $z = \zeta(\bar{s}_i, s_{-i})$. Let $z' \equiv \zeta(s_i, s_{-i})$, so $s_i \in S_i(z')$ and $s_{-i} \in S_{-i}(z')$. The claim implies that also $\bar{s}_i \in S_i(z')$. Then, by Remark 1, $(\bar{s}_i, s_{-i}) \in S(z')$, that is, $z' = \zeta(\bar{s}_i, s_{-i}) = z$, so $\bar{s}_i \in S_i(z)$ as well.

Part (iii) of the theorem has two implications, which will be used below.

Implication (iii.a): s_i is structurally rational given μ_i if and only if $\bar{s}_i$ is. *Proof*: By (3) and Remark 1, $(s_i, s_{-i}) \in S(z)$ iff $(\bar{s}_i, s_{-i}) \in S(z)$, so that $U_i(s_i, s_{-i}) = U_i(\bar{s}_i, s_{-i})$ for all $s_{-i} \in S_{-i}$ and, therefore, $U_i(s_i, p) = U_i(\bar{s}_i, p)$ for every $p \in \Delta(S_{-i})$, which implies the claim.

Implication (iii.b): $\hat{s}_i^*$ is structurally rational given μ_i^*. *Proof*: The first two terms in equation (22) for $U_i^*(s_i^*, q)$ and $U_i^*(\hat{s}_i^*, q)$ are the same, because $\mathbf{d}_i(s_i^*) = s_i = \mathbf{d}_i(\hat{s}_i^*)$ and $\mathbf{w}_i(s_i^*) = w_i = \mathbf{w}_i(\hat{s}_i^*)$, By (3), the third term is also the same, because $\mathbf{d}_i(s_i^*)$ and $\mathbf{r}_i(s_i^*)$ are realization-equivalent, and by construction $\mathbf{d}_i(\hat{s}_i^*) = s_i = \mathbf{r}_i(\hat{s}_i^*)$. Hence, $U_i^*(\hat{s}_i^*, q) = U_i^*(s_i^*, q)$ for every $q \in \Delta(S_{-i}^*)$. Since s_i^* is structurally rational given μ_i^*, so is $\hat{s}_i^*$.

To prove part (i) of the theorem, by implication (iii.a), it is enough to show that s_i is structurally rational given μ_i. Since, by implication (iii.b), $\hat{s}_i^*$ is structurally rational given μ_i^*, there is a perturbation $(q^k)_{k\geq 1}$ of μ_i^* such that $U_i^*(\hat{s}_i^*, q^k) \geq U_i^*(t_i^*, q^k)$ for all k and $t_i^* \in S_i^*$. Fix $t_i \in S_i$ arbitrarily and let $t_i^* = s_i^*(t_i, w_i, t_i)$. Then by construction $\mathbf{r}_i(\hat{s}_i^*) = \mathbf{d}_i(\hat{s}_i^*) = s_i$, $\mathbf{r}_i(t_i^*) = \mathbf{d}_i(t_i^*) = t_i$, and $\mathbf{w}_i(\hat{s}_i^*) = \mathbf{w}_i(t_i^*)$. Therefore, for every $k \geq 1$, the second and third terms in equation (22) for $U_i^*(\hat{s}_i^*, q^k)$ and $U_i^*(t_i^*, q^k)$ have the same value, so

$$0 \leq U_i^*\big(\hat{s}_i^*, q^k\big) - U_i^*\big(t_i^*, q^k\big) = \frac{1}{3}\sum_{s_{-i}} q^k\big(\{i\}\times[s_{-i}]\big)\big[U_i(s_i, s_{-i}) - U_i(t_i, s_{-i})\big]. \tag{23}$$

Since μ_i^* agrees with μ_i, by Lemma 3 part (i), there $\kappa \geq 1$ such that $q^{k+\kappa-1}(\{i\}\times S_{-i0}^*) > 0$ for all $k \geq 1$ and the sequence $(p^k)_{k\geq 1}$ defined in equation (21) is a perturbation of μ_i. Then equation (23) implies that, for this perturbation, $U_i(s_i, p^k) \geq U_i(t_i, p^k)$ for all k. Since t_i was arbitrary, s_i is structurally rational for μ_i.

For part (ii) of the theorem, suppose $w_i = p$, and let $\hat{s}_i^* = s_i^*(s_i, p, s_i)$. By contradiction, suppose that $\mu_i(E|I_i) > p$, and let $t_i^* = s_i^*(s_i, E, s_i)$. I show that, for all perturbations (q^k) of μ_i^*, eventually $U_i^*(\hat{s}_i^*, q^k) < U_i^*(t_i^*, q^k)$, which contradicts the fact that $\hat{s}_i^*$ is structurally rational by implication (iii.b).

Since by construction $\mathbf{r}_i(\hat{s}_i^*) = \mathbf{r}_i(t_i^*) = s_i$, $\mathbf{d}_i(\hat{s}_i^*) = \mathbf{d}_i(t_i^*) = s_i$, $\mathbf{w}_i(s_i^*) = p$ and $\mathbf{w}_i(t_i^*) = E$, for every $q \in \Delta(S_{-i}^*)$ the first and third terms in equation (22) for $U_i^*(s_i^*, q)$ and $U_i^*(t_i^*, q)$ are equal, and

$$\begin{aligned} U_i^*\big(\hat{s}_i^*, q\big) - U_i^*\big(t_i^*, q\big) &= \frac{1}{3}\sum_{s_{-i}\in S_{-i}} q\big(\{i\}\times[s_{-i}]\big)\big[B_i(p, s_{-i}) - B_i(E, s_{-i})\big] \\ &= \frac{1}{3}\Bigg[\sum_{s_{-i}\in E} q\big(\{i\}\times[s_{-i}]\big)(p-1) + \sum_{s_{-i}\in S_{-i}(I_i)\setminus E} q\big(\{i\}\times[s_{-i}]\big)p\Bigg] \end{aligned}$$

$$= \frac{1}{3}\left[p \sum_{s_{-i}\in S_{-i}(I_i)} q\big(\{i\}\times[s_{-i}]\big) - \sum_{s_{-i}\in E} q\big(\{i\}\times[s_{-i}]\big)\right]$$

$$= \frac{1}{3}\left[p\cdot q\big(S^*_{-i}(I_{s_i,w_i})\big) - q\Big(\{i\}\times\bigcup\big\{[s_{-i}]:s_{-i}\in E\big\}\Big)\right], \qquad (24)$$

where the last equality follows from equation (19).

Since by assumption $\mu_i(E|I) > p$ and μ^*_i agrees with μ_i, $\mu^*_i(\{i\}\times\bigcup\{[s_{-i}]:s_{-i}\in E\}|I_{\bar{s}_i,p}) = \mu^*_i(\{i\}\times\bigcup\{[s_{-i}]:s_{-i}\in E\}|I_{\bar{s}_i,E}) > p$. Hence, for any perturbation $\{q^k\}_{k\geq 1}$ of μ^*_i, and all $w_i\in W_i$,

$$p < \lim_{k\to\infty} q^k\Big(\{i\}\times\bigcup\Big\{[s_{-i}]:s_{-i}\in E\Big\}\Big|S^*_{-i}(I_{\bar{s}_i,w_i})\Big) = \lim_{k\to\infty}\frac{q^k\Big(\{i\}\times\bigcup\big\{[s_{-i}]:s_{-i}\in E\big\}\Big)}{q^k\big(S^*_{-i}(I_{\bar{s}_i,w_i})\big)};$$

the last equality uses the fact that, by equation (20), $E\subseteq S_{-i}(I)$ implies $\{i\}\times\bigcup\{[s_{-i}]:s_{-i}\in E\}\subseteq S^*_{-i}(I_{\bar{s}_i,w_i})$. Hence, for large k, $p\cdot q^k(S^*_{-i}(I_{\bar{s}_i,w_i})) - q^k(\{i\}\times\bigcup\{[s_{-i}]:s_{-i}\in E\})<0$, and by equation (24), $U^*_i(\hat{s}^*_i,q^k) < U^*_i(t^*_i,q^k)$, as claimed. The case $w_i = E$ is analogous, hence omitted.

Finally, to prove the last claim in Theorem 2, suppose that $s_i\in S_i$ is structurally rational given μ_i, so there is a perturbation $(p^k)_{k\geq1}$ of μ_i such that $U_i(s_i,p^k)\geq U_i(t_i,p^k)$ for all $k\geq 1$ and all $t_i\in S_i$. Moreover, if $W_i=\{E,p\}$, either $p^k(E|S_{-i}(I_i))\geq p$ infinitely often; otherwise, or $p^k(E|S_{-i}(I_i))\leq p$ eventually. In the former case, restrict attention to a subsequence of (p^k) for which $p^k(E|S_{-i}(I_i))\geq p$ and let $w_i=E$; in the latter, restrict attention to a subsequence of (p^k) for which $p^k(E|S_{-i}(I_i))\leq p$ and let $w_i=p$. In either case, to simplify indices, I abuse notation and refer to the resulting subsequence also as $(p^k)_{k\geq1}$. If instead $W_i=\{*\}$, then let $w_i=*$.

Let $s^*_i\equiv s^*_i(s_i,w_i,w_i)$, so $\mathbf{d}_i(s^*_i)=\mathbf{r}_i(s^*_i)=s_i$. Since μ^*_i agrees with μ_i, by Lemma 3, part (ii), there is a perturbation $(q^k)_{k\geq1}$ of μ^*_i that satisfies equation (21) with $\kappa=1$. It must be shown that $U^*_i(s^*_i,q^k)\geq U^*_i(t^*_i,q^k)$ for all $k\geq1$ and $t^*_i\in S^*_i$. Fix one such t^*_i arbitrarily.

Since $q^k(\{i\}\times S^*_{-0i})>0$ eventually as $(q^k)_{k\geq1}$ is a perturbation of μ^*_i, for all $r_i\in S_i$, eventually

$$U_i\big(r_i,p^k(\{s_{-i}\})\big) = \sum_{s_{-i}} q^k\big(\{i\}\times[s_{-i}]\big|\{i\}\times S^*_{-0i}\big)\cdot U_i(r_i,s_{-i})$$

$$= \frac{1}{q^k\big(\{i\}\times S_{-0i}\big)}\sum_{s_{-i}} q^k\big(\{i\}\times[s_{-i}]\big)\cdot U_i(r_i,s_{-i}).$$

Therefore, in particular, there is $K\geq1$ such that, for all $k\geq K$,[18]

$$\sum_{s_{-i}} q^k\big(\{i\}\times[s_{-i}]\big)\cdot U_i\big(\mathbf{d}_i(s^*_i),s_{-i}\big) \geq \sum_{s_{-i}} q^k\big(\{i\}\times[s_{-i}]\big)\cdot U_i\big(\mathbf{d}_i(t^*_i),s_{-i}\big).$$

Furthermore, since $\mathbf{d}_i(s^*_i)=\mathbf{r}_i(s^*_i)$, for all k,

$$\sum_{\substack{s_{-i}\in S_{-i}:\\ \zeta(\mathbf{d}_i(s^*_i),s_{-i})=\zeta(\mathbf{r}_i(s^*_i),s_{-i})}} q^k\big(\{i\}\times[s_{-i}]\big) = 1 \geq \sum_{\substack{s_{-i}\in S_{-i}:\\ \zeta(\mathbf{d}_i(t^*_i),s_{-i})=\zeta(\mathbf{r}_i(t^*_i),s_{-i})}} q^k\big(\{i\}\times[s_{-i}]\big).$$

[18]The choice of K is only to ensure that $q^k(\{i\}\times S^*_{-0i})>0$.

Finally, if $\mathbf{w}_i(s_i^*) = \mathbf{w}_i(t_i^*)$, then for all k,

$$\sum_{s_{-i}\in S_{-i}} q^k\big(\{i\}\times[s_{-i}]\big)\cdot B_i\big(\mathbf{w}_i\big(s_i^*\big), s_{-i}\big) = \sum_{s_{-i}\in S_{-i}} q^k\big(\{i\}\times[s_{-i}]\big)\cdot B_i\big(\mathbf{w}_i\big(t_i^*\big), s_{-i}\big).$$

Otherwise, necessarily $W_i = \{E, p\}$. Suppose $\mathbf{w}_i(s_i^*) = w_i = E$, so $\mathbf{w}_i(t_i^*) = p$. Since we restricted attention to a subsequence for which $p^k(E|S_{-i}(I)) \geq p$ or $p^k(E) \geq p \cdot p^k(S_{-i}(I))$,

$$\begin{aligned}
&\sum_{s_{-i}\in S_{-i}} q^k\big(\{i\}\times[s_{-i}]\big)\cdot B_i\big(\mathbf{w}_i\big(s_i^*\big), s_{-i}\big) \\
&\quad = \sum_{s_{-i}\in E} q^k\big(\{i\}\times[s_{-i}]\big)\cdot 1 = p^k(E)\cdot q^k\big(\{i\}\times S_{-0i}\big) \\
&\quad \geq p\cdot p^k\big(S_{-i}(I)\big)\cdot q^k\big(\{i\}\times S_{-0i}\big) = \sum_{s_{-i}\in S_{-i}(I)} q^k\big(\{i\}\times[s_{-i}]\big)\cdot p \\
&\quad = \sum_{s_{-i}\in S_{-i}} q^k\big(\{i\}\times[s_{-i}]\big)\cdot B_i\big(\mathbf{w}_i\big(t_i^*\big), s_{-i}\big),
\end{aligned}$$

where the second and third equalities follow from equation (21). Similarly, if $\mathbf{w}_i(s_i^*) = w_i = p$, then $\mathbf{w}_i(t_i^*) = E$ and $p^k(E|S_{-i}(I)) \leq p$ and an analogous argument yields the same inequality.

Therefore, equation (22) implies that, for all $k \geq K$, $U_i^*(s_i^*, q^k) \geq U_i^*(t_i^*, q^k)$. Since t_i^* was arbitrary, s_i^* is structurally rational for μ_i^*. This completes the proof of Theorem 2. *Q.E.D.*

REFERENCES

AUMANN, ROBERT J., AND JACQUES H. DREZE (2009): "Assessing Strategic Risk," *American Economic Journal: Microeconomics*, 1 (1), 1–16. [2437]

BATTIGALLI, PIERPAOLO (1997): "On Rationalizability in Extensive Games," *Journal of Economic Theory*, 74 (1), 40–61. [2438]

BATTIGALLI, PIERPAOLO, AND MARCIANO SINISCALCHI (2002): "Strong Belief and Forward Induction Reasoning," *Journal of Economic Theory*, 106 (2), 356–391. [2438,2440]

BECKER, GORDON M., MORRIS H. DEGROOT, AND JACOB MARSCHAK (1964): "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, 9 (3). [2443]

BEN-PORATH, ELCHANAN (1997): "Rationality, Nash Equilibrium and Backwards Induction in Perfect-Information Games," *The Review of Economic Studies*, 23–46. [2440,2449]

BEWLEY, TRUMAN (2002): "Knightian Decision Theory: Part I," *Decisions in Economics and Finance*, 25 (2), 79–110. (first version 1986). [2438,2442]

BLANCO, MARIANA, DIRK ENGELMANN, ALEXANDER K. KOCH, AND HANS-THEO NORMANN (2010): "Belief Elicitation in Experiments: Is There a Hedging Problem?" *Experimental Economics*, 13 (4), 412–438. [2437]

COSTA-GOMES, MIGUEL A., AND GEORG WEIZSÄCKER (2008): "Stated Beliefs and Play in Normal-Form Games," *The Review of Economic Studies*, 75 (3), 729–762. [2437]

GILBOA, ITZHAK, AND DAVID SCHMEIDLER (2003): "A Derivation of Expected Utility Maximization in the Context of a Game," *Games and Economic Behavior*, 44 (1), 172–182. [2437]

GOVINDAN, SRIHARI, AND ROBERT WILSON (2009): "On Forward Induction," *Econometrica*, 77 (1), 1–28. ISSN 1468-0262. [2450]

KOHLBERG, ELON, AND PHILIP J. RENY (1997): "Independence on Relative Probability Spaces and Consistent Assessments in Game Trees," *Journal of Economic Theory*, 75 (2), 280–313. [2440]

KREPS, DAVID M., AND ROBERT WILSON (1982): "Sequential Equilibria," *Econometrica: Journal of the Econometric Society*, 50 (4), 863–894. [2441,2443]

LUCE, R. DUNCAN, AND HOWARD RAIFFA (1957): *Games and Decisions*. New York: Wiley. [2437]

MAILATH, GEORGE J., LARRY SAMUELSON, AND JEROEN M. SWINKELS (1993): "Extensive Form Reasoning in Normal Form Games," *Econometrica*, 61, 273–302. [2440]
MYERSON, ROGER B. (1986): "Multistage Games With Communication," *Econometrica*, 54 (2), 323–358. ISSN 0012-9682. [2440]
NYARKO, YAW, AND ANDREW SCHOTTER (2002): "An Experimental Study of Belief Learning Using Elicited Beliefs," *Econometrica*, 70 (3), 971–1005. [2437]
OSBORNE, MARTIN J., AND ARIEL RUBINSTEIN (1994): *A Course on Game Theory*. Cambridge, MA: MIT Press. [2439,2440,2445,2451]
RENY, PHILIP J. (1992): "Backward Induction, Normal Form Perfection and Explicable Equilibria," *Econometrica*, 60 (3), 627–649. ISSN 0012-9682. [2437,2438,2441,2443]
RÉNYI, ALFRED (1955): "On a New Axiomatic Theory of Probability," *Acta Mathematica Hungarica*, 6 (3), 285–335. ISSN 0236-5294. [2440]
REY-BIEL, PEDRO (2009): "Equilibrium Play and Best Response to (Stated) Beliefs in Normal Form Games," *Games and Economic Behavior*, 65 (2), 572–585. [2437]
ROCKAFELLAR, R. TYRRELL (1970): *Convex Analysis*. Princeton, New Jersey: Princeton University Press. [2452]
RUBINSTEIN, ARIEL (1991): "Comments on the Interpretation of Game Theory," *Econometrica*, 59 (4), 909–924. ISSN 0012-9682. [2441]
SAVAGE, LEONARD J. (1954): *The Foundations of Statistics*. New York: Wiley. [2437]
SELTEN, REINHARD (1967): "Ein Oligopolexperiment mit Preisvariation und Investition," in *Beiträge zur experimentellen Wirtschaftsforschung*, ed. by H. Sauermann. Tübingen: JCB Mohr (Paul Siebeck), 103–135. [2437-2439]
——— (1975): "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4 (1), 25–55. ISSN 0020-7276. [2438,2450]
SINISCALCHI, MARCIANO (2022): "Foundations for Sequential Preferences," Report, Northwestern University. [2439,2440,2442,2449]
——— : "Putting Structural Rationality to Work," Report, in preparation, Northwestern University. [2439, 2450]
VAN DAMME, ERIC (1989): "Stable Equilibria and Forward Induction," *Journal of Economic Theory*, 48 (2), 476–496. [2437,2449]

Co-editor Barton L. Lipman handled this manuscript.

Manuscript received 11 September, 2018; final version accepted 16 June, 2022; available online 23 June, 2022.