Structural Rationality in Dynamic Games

Marciano Siniscalchi

September 30, 2021

Abstract

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 assumption that players are sequentially rational, 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 beliefs concerning off-path actions, as well as off-path beliefs, 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.

Economics Department, Northwestern University, Evanston, IL 60208; marciano@northwestern.edu.
Earlier drafts were circulated with the titles 'Behavioral counterfactuals,' 'A revealed-preference theory of strategic counterfactuals,' 'A revealed-preference theory of sequential rationality,' and 'Sequential preferences and sequential rationality.' I thank Bart Lipman and 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 RUD 2011, D-TEA 2013, and many seminar presentations for helpful comments on earlier drafts.

1 Introduction

Solution concepts for dynamic games, such as subgame-perfect, sequential, or perfect Bayesian equilibrium, aim to ensure that on-path play is sustained by "credible threats:" players believe that the (optimal) continuation play following any deviation from the predicted path would lead to a lower payoff. A credible threat involves two types of assumptions about beliefs. The first pertains to on-path beliefs about off-path play: what is the threat? The second pertains to beliefs at off-path information sets about subsequent play: why is the threatened course of action credible? What is it a best reply to? The assumptions placed on such beliefs are possibly the most important dimension in which solution concepts differ.

A key conceptual aspect of Savage (1954)'s foundational analysis of expected utility (EU) is to argue that the psychological notion of "belief" can and should be related to observable behavior. The objective of this paper is to characterize the behavioral content of assumptions on the beliefs players hold at any information set, whether on or off the predicted path of play. The results in this paper strengthen the foundations of dynamic game theory, and broaden the range of predictions that can be tested experimentally.

In a single-person decision problem, the individual's beliefs can be elicited by offering her "side bets" on the relevant uncertain 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 by offering side bets on her opponents' actions (Luce and Raiffa, 1957, §13.6); for game-theoretic experiments implementing side bets, see, e.g., Nyarko and Schotter (2002), Costa-Gomes and Weizsäcker (2008), Rey-Biel (2009), and Blanco, Engelmann, Koch, and Normann (2010).[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), the profile $(Out,(S,S))$ is a subgame-perfect equilibrium: Ann chooses *Out* at the initial

[1] For related approaches, see Aumann and Dreze (2009) and Gilboa and Schmeidler (2003).

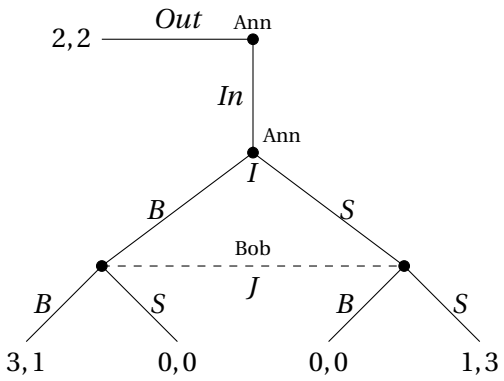


Figure 1: The Battle of the Sexes with an Outside Option

node under the threat that the Nash profile (S,S) would prevail in the subgame following *In*. 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 vs. 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 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 Ann initially expects Bob to play S in the subgame. It would appear that it would be enough to offer Ann a side bet on Bob's move. 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 sum up, elicitation must be carried out at the beginning of the game. However, under the textbook expected-utility assumption, beliefs at off-path information sets cannot be de-

rived by updating prior beliefs. Furthermore, a player's on-path beliefs cannot be elicited in an incentive-compatible way if that player's own moves can cause certain information sets not to be reached.

To address these issues, I propose the notion of *structural rationality*, which combines insights from Selten (1975) and Bewley (2002). Player i's strategy s_i is deemed superior to another strategy t_i if it yields a higher (ex-ante) expected payoff against all feasible perturbations, or trembles, of i's beliefs—that is, all perturbations that assign positive probability to every information set of i, and approximate i's conditional beliefs there. Thus, structural rationality reflects the player's *ex-ante* perspective—the relevant one when bets are offered at the beginning of the game. Moreover, it displays *informational caution*: the player takes into account the information she may possibly receive in the course of game play, even if some such information has zero probability ex-ante. Finally, it incorporates *robustness*: by considering all possible feasible perturbations, a player takes into account the possibility of her beliefs being misspecified, but does not commit to any particular form of misspecification. In sum, structural rationality takes the *possibility* of surprises seriously, without committing to specific 'theories' about them.

Theorem 1 shows that structural rationality implies weak sequential rationality (Reny, 1992; Battigalli, 1997; Battigalli and Siniscalchi, 2002) Theorem 2 provides a partial converse: if a strategy is weakly sequentially for a player given her beliefs, and there is no "relevant tie" for that player, (Battigalli, 1997), then that strategy is also structurally rational.

The main result of this paper, Theorem 3, 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 indicate their own *intended strategies*, and are rewarded if their actual play conforms to their predictions. This can be viewed as a variant of the *strategy method* of Selten (1967): the latter requires that players commit to (rather than just announce) extensive-form strategies. Struc-

tural rationality ensures that players have strict incentives to report the strategy that they are in fact planning to follow. This enables the elicitation of beliefs at all information sets, whether on-path or off-path. Assuming structural, rather than sequential rationality is essential to this result: see Example 4.

The companion paper Siniscalchi (2020a) provides an axiomatic behavioral analysis of structural rationality. Siniscalchi (2020a) also indicates that structural preferences are the coarsest (i.e., minimal) relation that still allows the behavioral identification of beliefs and utilities (see Theorem 2 in Siniscalchi, 2020a). Taken together, the present paper and Siniscalchi (2020a) establish foundations for dynamic game theory that are comparable to those provided by Luce and Raiffa (1957) for games with simultaneous moves.

A second companion paper, Siniscalchi (2021), provides an alternative, computationally efficient characterization of structural rationality, and demonstrates how to incorporate it into equilibrium and non-equilibrium solution concepts. Section 6.F in the present paper takes a first step and defines a version of sequential equilibrium in which structural rationality is the behavioral hypothesis. It also draws a connection with trembling-hand perfect equilibrium.

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, Def. 200.1, pp. 200-201), a finite dynamic game with imperfect information is represented by a tuple $\left(N, A, Z, P, (\mathscr{I}_i, u_i)_{i\in N}\right)$, where:

- N is the set of **players** and A is the set of **actions**.
- $Z \subset \bigcup_{0\leq t<\infty} A^t$ is the finite set of **terminal histories**. Given Z, $H \equiv \bigcup_{(a^1,\ldots,a^t)\in Z}\{(a^1,\ldots,a^\tau): 0\leq\tau\leq t\}$ is the set of all **histories**, including the **root** (empty history) ϕ.
- $P: H\setminus Z \to N$ is the **player function**.

- $\mathscr{I}_i$ is the collection of **information sets** of player i; it is a partition of $P^{-1}(\{i\})$, and is such that, if $(a_1,\ldots,a_K),(b_1,\ldots,b_L) \in I$ for some $I \in \mathscr{I}_i$, and $(a_1,\ldots,a_K,a) \in H$, then $(b_1,\ldots,b_L,a) \in H$. 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 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 $I \in \mathscr{I}_i$, $A(I) = \{a \in A : \exists (a_1,\ldots,a_k) \in I, (a_1,\ldots,a_k,a) \in H\}$ is the (non-empty) set of **actions available to** i **at** I.[2]
- For every $i \in N$, $S_i = \prod_{I \in \mathscr{I}_i} A(I)$ is the set of **strategies** of player i; the action specified by $s_i \in S_i$ at $I \in \mathscr{I}_i$ 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 $h = (a_1,\ldots,a_K) \in H$, $S(h) = \left\{ s \in S : \forall k = 1,\ldots,K, \exists i \in N, I \in \mathscr{I}_i \text{ s.t. } (a_1,\ldots,a_{k-1}) \in I, a^k = s_i(I) \right\}$ is the set of strategy profiles that **induce** h. Let $S_i(h) = \text{proj}_{S_i} S(h)$ and $S_{-i}(h) = \text{proj}_{S_{-i}} S(h)$.
- For every $i \in N$ and $I \in \mathscr{I}_i$, $S(I) = \bigcup_{h \in I} S(h)$ 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.[3]
- 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 $(s_i, s_{-i}) \in S(z)$. For any probability $p \in \Delta(S_{-i})$ and $s_i \in S_i$, i's expected payoff given p if she plays s_i is $U_i(s_i, p) = \sum_{s_{-i}} U_i(s_i, s_{-i}) \cdot p(\{s_{-i}\})$.

Sets of the form $S_{-i}(I)$, for $I \in \mathscr{I}_i$, are called **conditioning events**.

I assume that the game has **perfect recall**, analogously to Def. 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 $I \in \mathscr{I}_i$, $S(I) = S_i(I) \times S_{-i}(I)$. Second, the set $S(I)$ satisfies **strategic independence** (Mailath, Samuelson, and Swinkels, 1993, Definition 2 and Theorem 1): for every

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

[3] That is: if i's co-players follow the profile s_{-i}, I *can* be reached; whether it *is* reached depends upon i's play.

$s_i, t_i \in S_i(I)$ 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 $\left(N, A, Z, P, (\mathscr{I}_i, u_i)_{i \in N}\right)$.

I represent player i's beliefs as a collection of probability distributions over co-players' strategies, indexed by her information sets $I \in \mathscr{I}_i$:[4] cf. 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 $(\mu(\cdot|I))_{I \in \mathscr{I}_i \cup \{\phi\}}$ 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 $(\mu(\cdot|I))_{I \in \mathscr{I}_i \cup \{\phi\}}$ 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 *A **consistent conditional probability system (CCPS)** for player i is an array $\mu = \left(\mu(\cdot|I)\right)_{I \in \mathscr{I}_i \cup \{\phi\}} \in \Delta(S_{-i})^{\mathscr{I}_i \cup \{\phi\}}$ for which there exists a sequence $(p^k)_{k \geq 1} \in \Delta(S_{-i})^{\mathbf{N}}$, called a **perturbation** of μ, such that, for all $I \in \mathscr{I}_i \cup \{\phi\}$, $p^k(S_{-i}(I)) > 0$ for all $k \geq 1$, and $\lim_{k \to \infty} p^k(\cdot|S_{-i}(I)) = \mu(\cdot|I)$. Denote the set of CCPSs for player i by $\Delta(S_{-i}, \mathscr{I}_i)$.*

[4] Definition 1 implies that, equivalently, one can take the corresponding conditioning events $S_{-i}(I)$ as indices.

A CCPS is a "conditional probability system" in the sense of Rényi (1955): this follows from minor modifications of arguments in Myerson (1986). However, it satisfies additional restrictions: see Siniscalchi (2021).

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 $\mu = \mu(\cdot|\phi)$.

Finally, I formalize the notion of sequential rationality used in this paper. Following Reny (1992) and Rubinstein (1991), the definition I adopt does not restrict the actions specified by a strategy s_i of player i at information sets that s_i does not allow. As these authors have argued, such restrictions may be interpreted as (equilibrium) assumptions on the beliefs of i's co-players which reflect the logic of backward induction: they do not characterize player i's rational decision-making. I follow Reny (1992) and call the resulting notion "weak sequential rationality," to distinguish it from the definition in Kreps and Wilson (1982).

Definition 2 *Fix a CCPS $\mu \in \Delta(S_{-i}, \mathscr{I}_i)$. A strategy $s_i \in S_i$ is **weakly sequentially rational given** μ if, for every $I \in \mathscr{I}_i \cup \{\phi\}$ 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

It is now possible to formally define structural rationality. For conciseness, all definitions and results in this section apply to a player $i \in N$, and a CCPS $\mu \in \Delta(S_{-i}, \mathscr{I}_i)$ for player i in the dynamic game $\left(N, A, Z, P, (\mathscr{I}_i, u_i)_{i \in N}\right)$.

Definition 3 *For all strategies $s_i, t_i \in S_i$, t_i is **structurally strictly preferred to** s_i **given** μ, written $t_i \succ^{\mu} s_i$, if $U_i(t_i, p^k) > U_i(s_i, p^k)$ eventually[5] for all perturbations $(p^k)_{k \geq 1}$ of μ. A strategy s_i is **structurally rational given** μ if there is no $t_i \in S_i$ with $t_i \succ^{\mu} s_i$.*

[5] That is, for all sufficiently large k.

Definition 3 is in the spirit of Bewley (2002)'s representation of ambiguity, or Knightian uncertainty (Ellsberg, 1961). Expected payoffs are computed with respect to perturbations, rather than individual probabilities: intuitively, the structurally rational agent perceives ambiguity about "infinitesimal" deviations from her CCPS.

Remark 1 *Strategy s_i is structurally rational given μ if and only if, for every $t_i \in S_i$, 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 k.*

The perturbations in Remark 1 may depend upon the specific strategy t_i that is being compared to s_i: an example can be found in the previous version of this paper, Siniscalchi (2020b).

Structural rationality depends upon (i) the extensive-form structure of the game, and specifically on the collection $\{S_{-i}(I) : I \in \mathscr{I}_i \cup \{\phi\}\}$ 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 $I \in \mathscr{I}_i$. By Remark 1, *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 $(Out, (S,S))$: that is, $\mu(\{Out\}|\phi) = 1$ and $\mu(\{InS\}|J) = 1$. Then any perturbation $(p^k)_{k\geq 1}$ of μ must assign positive probability to $S_{-b}(J) = \{InS, InB\}$, and furthermore $p^k(\{InS\}|\{InS, InB\}) \to 1$. Thus, for k large enough, $U_b(S, p^k) > U_b(B, p^k)$: therefore, $S \succ^{\mu} B$. □

Example 2 The game in Figure 2 is parameterized by $x \in [0,2]$; for $x < 2$, it is a Centipede game. Ann's beliefs μ satisfy $\mu(\{d\}|\phi) = \mu(\{a\}|I) = 1$.

Denote by D_1 either one of the realization-equivalent strategies D_1D_2, D_1A_2. If $x < 2$, then D_1 is the unique structurally rational strategy given μ: any perturbation $(p^k)_{k\geq 1}$ of μ must assign positive probability to $\{a\} = S_b(I)$, but also satisfy $p^k(\{d\}) = p^k(\{d\}|S_b(\phi)) \to 1$, so even-

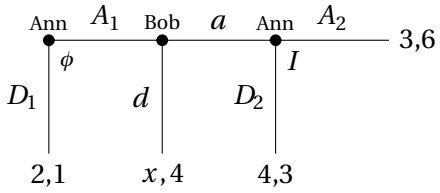


Figure 2: A centipede-like game.

tually $U_a(D_1, p^k) > U_a(s_a, p^k)$ for any other strategy $s_a \neq D_1$ of Ann. Of course, D_1 is also the unique weakly sequential best reply to μ. In addition, since $p^k(\{a\}) > 0$ for all perturbations and all k, it is also the case that $A_1D_2 \succ^\mu A_1A_2$. Thus, structural preferences do not just yield a maximal (i.e., structurally rational) collection of strategies: they can also rank inferior strategies.

For $x = 2$, A_1D_2 is the unique structurally rational strategy given μ: now $U_a(s_a, d) = 2$ for all $s_a \in S_a$, but since $p^k(\{a\}) > 0$ and $U_a(A_1D_2, a) = 4 > U_a(s_1, a)$ for all $s_a \neq A_1D_1$, $U_a(A_1D_2, p^k) > U_a(s_a, p^k)$ for all k and all other strategies $s_a \neq A_1D_1$. By comparison, both D_1 and A_1D_2 are weakly sequentially rational given μ. Thus, structural rationality sometimes refines weak sequential rationality, though only in the presence of ties: see Section 5.1. □

Example 3 The game in Figure 3 is an extension of "Matching Pennies" in which Bob has an additional choice, o, following which Ann may move again. 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 $U_a(HL, p^k) > U_a(HR, p^k)$, HR is not structurally rational given μ. But how about HL and T? If $2p^k(\{o\}) + p^k(\{h\}) > -p^k(\{o\}) + p^k(\{t\})$, then $U_a(HL, p^k) > U_a(T, p^k)$; for example, let $p^k(\{o\}) = \frac{1}{k}$ and $p^k(\{h\}) = p^k(\{t\}) = \frac{1}{2} - \frac{1}{2k}$. If however $2p^k(\{o\}) + p^k(\{h\}) < -p^k(\{o\}) + p^k(\{t\})$, then $U_a(HL, p^k) < U_a(T, p^k)$; for instance, let $p^k(\{o\}) = \frac{1}{8k}$, $p^k(\{h\}) = \frac{1}{2} - \frac{1}{2k}$,

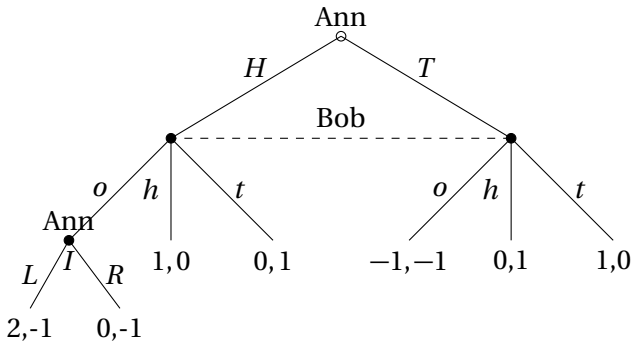


Figure 3: Modified Matching Pennies

and $p^k(\{t\}) = \frac{1}{2} + \frac{3}{8k}$. Thus, neither $HL \succ^\mu T$ nor $T \succ^\mu HL$, so both HL and T are structurally rational given μ. Of course, these are also the weakly sequentially rational best replies to μ.

This example illustrates the *robustness* requirement in Definition 3. Different perturbations of Ann's beliefs μ select one or the other strategy. The fact that one must take all such perturbations into account implies that both strategies are deemed structurally rational. Notice also that, in this example, one does *not* obtain a complete ranking of strategies. □

5 Main Results

5.1 Structural and Weak Sequential Rationality

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

The converse to Theorem 1 does not hold in general: see Example 2. However, structural and weak sequential rationality are "generically" equivalent. A dynamic game has a **relevant tie for player** i (Battigalli, 1997) if there is an information set $I \in \mathscr{I}_i$, strategies $s_i, t_i \in S_i(I)$, a profile $s_{-i} \in S_{-i}(I)$, and terminal histories $z, z' \in Z$ such that $(s_i, s_{-i}) \in S(z)$, $(t_i, s_{-i}) \in S(z')$, $z \neq z'$, and $u_i(z) = u_i(z')$. That is: starting from I, when co-players play according to s_{-i},

strategies s_i and t_i lead to different terminal histories, but player i receives the same payoff at those histories. "Not having relevant ties" is a particularly simple form of genericity, which in particular does not depend upon any particular CCPS under consideration.

Theorem 2 *Fix a player $i \in N$ and a CCPS $\mu \in \Delta(S_{-i}, \mathscr{I}_i)$. If $s_i \in S_i$ is weakly sequentially rational given μ, and there is no relevant tie for i, then s_i is structurally rational given μ.*

For instance, in the game in Figure 3, while Ann's payoffs are not all distinct, there are no relevant ties for Ann, so one can conclude that every weakly sequentially rational strategy is also structurally rational, regardless of Ann's beliefs.

5.2 Eliciting Conditional Beliefs

This section proposes an elicitation mechanism under which structurally rational players have strict incentives to bet in accordance with their conditional beliefs. The mechanism is designed so that, in eliciting a player's beliefs, one does not alter the other players' strategic incentives. This distinguishes belief elicitation in games from elicitation in decision problems.

To illustrate the main ideas with a minimum of notation, throughout this section I only consider *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 $I_i \in \mathscr{I}_i$ (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 enrich the set of bets offered to players, or to adapt the approach introduced here to alternative mechanisms (e.g. Becker, DeGroot, and Marschak, 1964).

At a broad level, the elicitation mechanism consists of two phases.

- In the first phase, each player i simultaneously chooses a *bet* $w_i \in \{E_i, p_i\}$, and an *intended strategy* $\bar{s}_i \in S_i$. Simultaneously, the experimenter—player 0—randomly selects one of the players, henceforth called "the selected player."

- In the second phase, the selected player plays the original game with the experimenter, who faithfully implements the intended 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 *intended* strategies of the other players, $\bar{s}_{-i}$; and a *bonus* $\epsilon > 0$ if her direct play is consistent with her intended strategy $\bar{s}_i$.

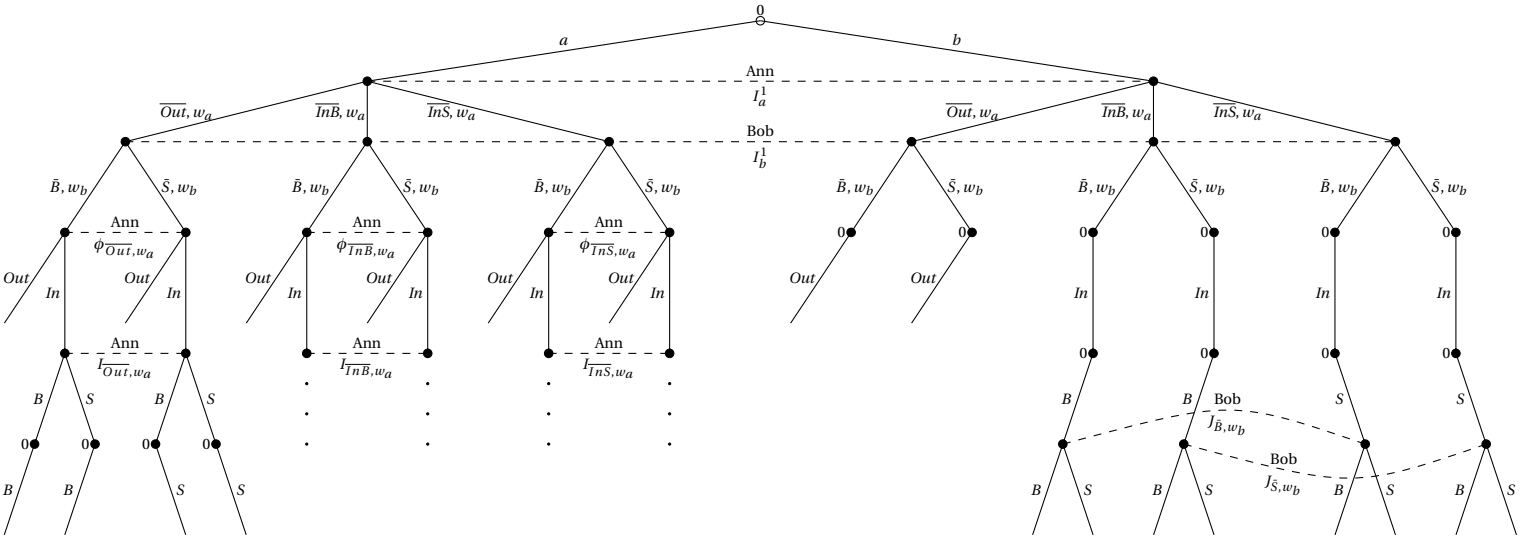


Figure 4: A stylized representation of the elicitation game tree for Figure 1

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 $(\bar{B}, p_b)$).

[6]Alternatively, players may play *separately* with the experimenter, either simultaneously or in sequence, provided they do not observe each other's moves. A version of Theorem 3 applies to these schemes as well, but formally describing such mechanisms is considerably more cumbersome.

I now formally define the elicitation game associated with an arbitrary dynamic game $\left(N, A, Z, P, (\mathscr{I}_i, u_i)_{i\in N}\right)$. 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 $i \in N$, $I_i \in \mathscr{I}_i$ 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

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

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 an intended 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 intended action. Formally, if the profile of intended 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 intended strategy $\bar{s}_n$, the selected player can choose any course of action that is also available in the original game. Thus,

$$Z^* = \left\{\left(n, (\bar{s}_1, w_1), \ldots, (\bar{s}_N, w_N), z\right) : n \in N, (\bar{s}_i, w_i) \in S_i \times W_i \,\forall\, i \in N, z \in Z, \bar{s}_{-n} \in S_{-i}(z)\right\} \tag{2}$$

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

As in Section 2, given the set Z^* of terminal histories, one can define the set H^* of all his-

[7] If $W_i = \{*\}$, the choice of I_i is immaterial.

tories, terminal or not. With this, the player function is defined as

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

Now turn to information. The experimenter has perfect information:

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

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 intended 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 intended strategy is either $\overline{InB}$ or $\overline{InS}$). 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 $I \in \mathscr{I}_i$ and $(\bar{s}_i, w_i) \in S_i \times W_i$, let

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

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

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

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

$$u_0^*(z^*) = 0 \tag{8}$$

$$u_i^*(z^*) = \begin{cases} 0 & n \neq i \\ \frac{1}{3}u_i(z) + \frac{1}{3}B(w_i, \bar{s}_{-i}) + \frac{1}{3}\cdot\epsilon\cdot 1_{\bar{s}_i \in S_i(z)} & \end{cases} \tag{9}$$

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

For the selected player i, $u_i(z)$ 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 intended 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 $\left(N^*, A^*, Z^*, P^*, (\mathscr{I}_i^*, u_i^*)_{i\in N\cup\{0\}}, \epsilon\right)$, where $\epsilon > 0$ and the other elements are as in Equations (1)–(9).*

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 playing directly, the selected player, say n, will choose a course of action that is structurally rational given the beliefs she holds at her information sets in the second phase of the mechanism.[9] But, fixing n's choice of an intended 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

[8] In particular, in Figure 4, if $\bar{s}_b = \bar{B}$, Bob is selected, and Ann chooses $\bar{s}_a = Out$, 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 intended strategy, he still receives the bonus ϵ,

[9] This is a loose statement because "structural rationality" is a property of the overall strategy of a player, not of its restriction to a part of the tree. This is made formal below, leveraging the separable structure of payoffs.

of action in 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 intended strategy $\bar{s}_n$. However, n *does* get a bonus if $\bar{s}_n$ is consistent with her direct play. This implies that, 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. Structural rationality implies that n will correctly anticipate not just her own on-path moves, but also moves following other players' unexpected actions. Moreover, by the previous argument, under belief agreement, her intended 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 *intended* 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 the above intuitive outline, the first step is to analyze the structure of strategies in the elicitation game. With a slight notational abuse, I identify the set of strategies S_0^* for the experimenter with N, the set of players, because at all other histories player 0 has a single available action. A strategy $s_i^* \in S_i^*$ for a player $i \in N$ must specify an intended strategy $\bar{s}_i$ and bet w_i at I_i^1. In addition, it must specify an action at *every* information set of the form $I_{\bar{t}_i, v_i}$ (see Equation (6)), even for pairs $(\bar{t}_i, v_i) \neq s_i^*(I_i^1) = (\bar{s}_i, w_i)$.[10] However, such actions can be disregarded as they are not payoff-relevant. To do so, define maps $\bar{\mathbf{s}}_i : S_i^* \to S_i$, $\mathbf{w}_i : S_i^* \to W_i$, and $\mathbf{s}_i : S_i^* \to S_i$ as follows: for every $s_i^* \in S_i^*$, if $s_i^*(I_i^1) = (\bar{s}_i, w_i)$ then $\bar{\mathbf{s}}_i(s_i^*) = \bar{s}_i$ and $\mathbf{w}_i(s_i^*) = w_i$; furthermore, given that $s_i^*(I_i^1) = (\bar{s}_i, w_i)$, the strategy $\mathbf{s}_i(s_i^*) \in S_i$ is defined by

$$\forall I \in \mathscr{I}_i, \qquad \mathbf{s}_i(s_i^*)(I) = s_i^*(I_{\bar{s}_i, w_i}). \tag{10}$$

With this notation in place, I can formalize the key assumption of belief agreement:

[10] Thus, the strategy set for player i is $S_i^* = (S_i \times W_i) \times \prod_{(\bar{s}_i, w_i) \in S_i \times W_i} \prod_{I \in \mathscr{I}_i} A(I) = (S_i \times W_i) \times S_i^{S_i \times W_i}$.

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

$$\mu^*\left(\left\{t^*_{-i} : t^*_0 = n, \bar{\mathbf{s}}_j(t^*_j) = s_j \ \forall j \in N \setminus \{i\}\right\} \middle| \phi^*\right) = \frac{1}{N}\mu(\{s_{-i}\}|\phi) \tag{11}$$

$$\mu^*\left(\left\{t^*_{-i} : t^*_0 = i, \bar{\mathbf{s}}_j(t^*_j) = s_j \ \forall j \in N \setminus \{i\}\right\} \middle| I_{\bar{s}_i, w_i}\right) = \mu(\{s_{-i}\}|I) \quad \forall I_{\bar{s}_i, w_i} \in \mathscr{I}^*_i. \tag{12}$$

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 co-players' choices of intended strategies; and (ii) at every information set, i holds the same beliefs about each co-player j's intended strategy as about j's strategy in the original game. See §6.C for further discussion.

Given $\mu \in \Delta(S_{-i}, \mathscr{I}_i)$, a CCPS $\mu^* \in \Delta(S^*_{-i}, \mathscr{I}^*_i)$ that agrees with μ always exists (this is established in Theorem 3). On the other hand, more than one CCPS for player i in the elicitation game may agree with her CCPS in the original game. This is because i may assign different probabilities to her co-players' choices of side bets in the elicitation game. However, these differences do not affect i's payoff, and so are irrelevant for her strategic reasoning.

The main result of this section can now be stated: if belief elicitation is implemented as described above, and players' beliefs about others' intended strategies are the same as in the original game, then (1) players' structural preferences over second-stage strategies are also unchanged, and (2) belief bounds can be elicited from initial, observable betting choices. Furthermore, (3) regardless of a player's beliefs, structural rationality implies that intended strategies should be consistent with observed play.

Theorem 3 *Fix a questionnaire $(I_i, W_i)_{i \in N}$ and let $\left(N^*, (S^*_i, \mathscr{I}^*_i, U^*_i)_{i \in N^*}, S^*(\cdot)\right)$ be the associated elicitation game. Fix a CCPS $\mu_i \in \Delta(S_{-i}, \mathscr{I}_i)$ for player $i \in N$. Then there exists a CCPS $\mu^*_i \in \Delta(S^*_{-i}, \mathscr{I}^*_i)$ that agrees with μ_i. For any such CCPS μ^*_i, and for all $s^*_i, t^*_i \in S^*_i$:*

(1) if $\bar{\mathbf{s}}_i(s^*_i) = \mathbf{s}_i(s^*_i)$, $\bar{\mathbf{s}}_i(t^*_i) = \mathbf{s}_i(t^*_i)$, and $\mathbf{w}_i(s^*_i) = \mathbf{w}_i(t^*_i)$,

$$s^*_i \succ^{\mu^*_i} t^*_i \quad \Leftrightarrow \quad \mathbf{s}_i(s^*_i) \succ^{\mu_i} \mathbf{s}_i(t^*_i).$$

(2) if $W_i = (E, p)$ *and* $\bar{\mathbf{s}}_i(s_i^*) = \bar{\mathbf{s}}_i(t_i^*)$, $\mathbf{s}_i(s_i^*) = \mathbf{s}_i(t_i^*)$, $\mathbf{w}_i(s_i^*) = p$ *and* $\mathbf{w}_i(t_i^*) = E$,

$$p > \mu_i(E|I_i) \quad \Rightarrow \quad s_i^* \succ^{\mu_i^*} t_i^* \qquad \text{and} \qquad p < \mu_i(E|I_i) \quad \Rightarrow \quad t_i^* \succ^{\mu_i^*} s_i^*.$$

Furthermore,

(3) if there is $z \in Z$ *such that* $\mathbf{s}_i(s_i^*) \in S_i(z)$ *but* $\bar{\mathbf{s}}_i(s_i^*) \notin S_i(z)$*, then any* $t_i^* \in S_i^*$ *such that* $\mathbf{s}_i(t_i^*) = \bar{\mathbf{s}}_i(t_i^*) = \mathbf{s}_i(s_i^*)$ *and* $\mathbf{w}_i(t_i^*) = \mathbf{w}_i(s_i^*)$ *satisfies* $t_i^* \succ^{\nu_i^*} s_i^*$ *for all* $\nu_i^* \in \Delta(S_{-i}^*, \mathscr{I}_i^*)$.

Hence, if $W_i = (E, p)$, s_i^* *is structurally rational given* μ_i^*, *and* $\mathbf{w}_i(s_i^*) = E$ *(resp.* $\mathbf{w}_i(s_i^*) = p$*), then* $\bar{\mathbf{s}}_i(s_i^*)$ *and* $\mathbf{s}_i(s_i^*)$ *are structurally rational given* μ_i, $\mu_i(E|I_i) \geq p$ *(resp.* $\mu_i(E|I_i) \leq p$*),*[11] *and for all* $z \in Z$, $\bar{\mathbf{s}}_i(s_i^*) \in S_i(z)$ *if and only if* $\mathbf{s}_i(s_i^*) \in S_i(z)$.

This result also provides 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 its behavioral hypothesis; for instance the "structural" version of sequential equilibrium defined in Section 6.F. Then, if indeed players conform to such a solution concept, the version of the strategy method proposed here will elicit their intended behavior.

Corollary 1 *Suppose that* $W_i = \{*\}$ *for all* $i \in N$. *Then, for all* $i \in N$ *and all* s_i^*, t_i^* *such that* $\bar{\mathbf{s}}_i(s_i^*) = \mathbf{s}_i(s_i^*)$ *and* $\bar{\mathbf{s}}_i(t_i^*) = \mathbf{s}_i(t_i^*)$, $s_i^* \succ^{\mu_i^*} t_i^*$ *if and only if* $\mathbf{s}_i(s_i^*) \succ^{\mu_i} \mathbf{s}_i(t_i^*)$. *In particular,* s_i^* *is structurally rational given* μ_i^* *in the elicitation game if and only if* $\mathbf{s}_i(s_i^*)$ *is structurally rational given* μ_i *in the original game.*

Theorem 3 depends crucially on the assumption that players are structurally rational. (Weak) sequential rationality is not sufficient to deliver this results, even if beliefs satisfy the agreement condition of Definition 6:

[11] A weak inequality is needed because, if $p = \mu_i(E|I)$, strategies $s_i^* \in [\bar{s}_i, E, s_i]$ and $t_i^* \in [\bar{s}_i, p, s_i]$ may be incomparable.

Example 4 Consider the game in Figure 1 and assume that $W_a = \{*\}$ and $W_b = \{\{InS\}, 0.5\}$, 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 $0 < \epsilon < 1$,[12] the following strategies are part of a sequential equilibrium. Ann plays $(\overline{Out}, *)$ 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 $J_{\bar{t}_b,\nu_b}$, for all $(\bar{t}_b, \nu_b) \in S_b \times W_b$. 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 intended strategy $(\bar{S}, 0.5)$; at I_b^1, Bob expects Ann to have chosen $(\overline{Out})$, and at each $J_{\bar{t}_b,\nu_b}$, he assigns probability one to Ann having chosen intended strategy $\overline{InS}$.

The key point is that Bob is asked to bet at the beginning of the game, and sequential rationality[13] only requires that he maximize his *ex-ante* expected payoff. Since in equilibrium Bob expects Ann to choose $\overline{Out}$, he expected the bet to be called off. Hence, he is indifferent between his betting choices. Under structural rationality, instead, Bob takes into account the possibility that Ann might deviate from $\overline{Out}$ at I_a^1 even when choosing his bet at I_b^1; since he assigns probability 1 to $\overline{InS}$ given $\{\overline{InS}, \overline{InB}\}$, he has a strict preference to bet on $\{\overline{InS}\}$. □

To reconcile Theorem 3 and Example 4 with the generic equivalence result described in Section 5.1, notice that elicitation games feature numerous relevant ties *by construction*. Specifically, take the perspective of Bob at I_b^1 in Figure 4. If Ann reports intended strategy $\overline{Out}$ at I, then for a fixed intended strategy $\bar{s}_b$, both of Bob's actions $(\bar{s}_b, \{InS\})$ and $(\bar{s}_b, 0.5)$ yield the same payoff, namely $2 + \epsilon$. This is a relevant tie. More generally, by construction, elicitation games have numerous relevant ties, and are such that structural rationality is strictly stronger than weak sequential rationality.

[12] The condition $\epsilon < 1$ is only needed to ensure that the direct-play payoff dominates in Ann's and Bob's decisions when they are selected and find themselves at an information set that their own prior moves prevent from reaching. In other words, it is required for sequential rationality, but not for weak sequential rationality.

[13] 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.A Incomplete-information games The analysis may also be adapted to accommodate incomplete information. Fix a dynamic game with N players, strategy sets S_i, terminal histories Z, and information sets $\mathscr{I}_i$ for each $i \in N$. Consider sets Θ_i of possible "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 $\mu_{\theta_i} \in \Delta(S_{-i} \times \Theta)^{\{\phi\} \cup \mathscr{I}_i}$; 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 $I \in \{\phi\} \cup \mathscr{I}_i$. If the sets Θ_j are finite, Definitions 1, 2, and 3 can be applied to each type $\theta_i \in \Theta_i$ separately; Theorems 1, 2 and 3 then have straightforward extensions. Otherwise, it is more convenient to take the characterization in Siniscalchi (2021) as the definition of structural preferences, in which case, again, the remaining results go through unmodified.

6.B Higher-order beliefs The proposed approach can also be adapted to elicit higher-order beliefs. Consider a two-player game for simplicity. The analyst begins by eliciting Ann's first-order beliefs about Bob's strategies, as in Section 5.2. She then elicits Bob's second-order beliefs by offering him side bets on both Ann's strategies *and* on her first-order beliefs. The required formalism is that of games with incomplete information, taking Θ_i to be the set of all CCPSs for each player i. The incomplete-information extension of Theorem 3 ensures that second-order beliefs can be elicited in an incentive-compatible way. The argument extends to beliefs of higher orders.

6.C Elicitation: the notion of agreement Per Equation (9), Player i's payoff depends solely upon her co-players' intended strategies, not their direct play; this motivates defining agreement as a condition on player i's beliefs about the former, and not the latter. That said, by parts (1) and (3) of Theorem 3, if a co-player j is structurally rational, then—regardless of his

CCPS—his direct play will in fact be realization-equivalent to his intended strategy, as well as with his play in the original game. For this reason, one can alternatively formulate belief agreement as two separate conditions: one relating i's beliefs about j's play in the original game and her beliefs about j's direct play in the elicitation game, and one stating that, at every information set, i is certain that j's intended strategy is realization-equivalent to his direct play. However, Definition 6 is notationally simpler.

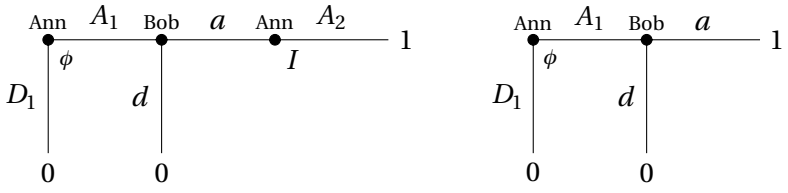
6.D Elicitation: modified or perturbed games In the equilibrium $(Out,(S,S))$ of the game of Figure 1, Ann's initial move prevents J from being reached. One might consider modifying the game so that J *is* actually reached, perhaps with small probability, regardless of Ann's initial move. However, such 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, e.g., Ben-Porath (1997), Example 3.2 and p. 36.

By way of contrast, the elicitation approach in Section 5.2 only modifies the game in ways that, as per Statement (1) of Theorem 3, are inessential for each player's structural preferences.

6.E Caution, elicitation, and triviality Consider the games in Figure 5a.[14] Ann has a single move available at I in Figure 5a. From the perspective of weak sequential rationality, such an information set can be disregarded. However, structural preferences treat the two games in

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

Figures 5a and 5b differently.



(a) I is trivial (b) I removed

Figure 5: A trivial information set; Ann's payoffs shown.

In Figure 5a, A_1A_2 is the only structurally rational strategy for Ann, if she assigns prior probability 1 to d. On the other hand, both D_1 and A_1A_2 are structurally rational, under the same prior belief, in the game in Figure 5b. The reason is that Ann has different conditioning events in the two games in Figure 5.

To avoid this, one can replace $\mathscr{I}_i$ with the collection $\mathscr{I}_i^{\text{nt}} = \{I \in \mathscr{I}_i : |A(I)| \geq 2\}$ of "non-trivial" information sets in Definitions 1, 2, 4 and 6: all 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' intended strategies in Definition 5.[15] The previous version of this paper, Siniscalchi (2020b), discusses other notions of "triviality" of information sets.

6.F Equilibrium and structurally rational strategies Siniscalchi (2021) incorporates structural rationality into solution concepts, including sequential equilibrium and extensive-form rationalizability. To provide a brief illustration, consider the former concept. Govindan and Wilson (2009) reformulate sequential equilibrium using (consistent) conditional probability systems; their definition is thus a convenient starting point. A *behavioral strategy* for player i

[15]This could be avoided by including moves by chance in the definition of a dynamic game. Doing so would entail some additional notational complexity, either in the definition of CCPSs (which would need to reflect the exogenous probability of chance moves) or in the way expected payoffs are calculated.

is an array $\beta = \big(\beta_i(I)\big)_{I\in\mathscr{I}_i} \in \Delta(A)^{\mathscr{I}_i}$ such that $\beta_i(I)(A(I)) = 1$ for all $I \in \mathscr{I}_i$. As usual, each behavioral strategy β_i induces a mixed strategy $\sigma_i \in \Delta(S_i)$; $\otimes_{j\neq i}\sigma_j$ 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, $\mu_i = \big(\mu_i(\cdot|I)\big)_{I\in\mathscr{I}} \in \Delta(S_{-i})^{\{\phi\}\cup\mathscr{I}_i}$, 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 $\big(\otimes_{j\neq i} p_j^k\big)(\cdot|S_{-i}(I)) \to \mu_i(\cdot|I)$ for each $I \in \mathscr{I}_i$.

(ii) For every i and $I \in \mathscr{I}_i$, 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 $I \in \mathscr{I}_i$, if $\beta_i(I)(a) > 0$, then there exists $s_i \in S_i(I)$ such that $s_i(I) = a$ and $t_i \succ^{\mu_i} s_i$ for *no* $t_i \in S_i(I)$ that satisfies $t_i(J) = s_i(J)$ for all $J \in \mathscr{I}_i$ that do not follow I.[16]

Refer to the companion paper Siniscalchi (2021) for an analysis of the resulting notion.

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 $p_{-i}^k \equiv \otimes_{j\neq i}\sigma_j^k$, $k \geq 1$. Each sequence $(p_{-i}^k)_{k\geq 1}$ defines a CCPS $\mu_{-i} \in \Delta(S_{-i}, \mathscr{I}_i)$ (possibly considering subsequences), and by Remark 1, every $s_i \in \operatorname{supp}\sigma_i$ is structurally rational given μ_{-i}.

[16] J follows I if, for every history $(a_1, \ldots, a_M) \in J$), there is $L < M$ such that $(a_1, \ldots, a_L) \in I$.

A Appendix: dynamic games

This section formalizes further properties of dynamic games, including perfect recall. It also introduces additional useful notation and results.

Fix a dynamic game $\left(N, A, Z, P, (\mathscr{I}_i, u_i)_{i \in N}\right)$ as defined in Section 2. Let H be the set of all (terminal and non-terminal) histories, as defined therein.[17]

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 $h, h' \in I \in \mathscr{I}_i$, then $X_i(h) = X_i(h')$. One immediate implication (used in the proof of Remark 2) 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 $s \in S(z)$.

This remark points out that the sets $S(h)$ (where h can be terminal or not) have a product structure.

Remark 2 *Let $h = (a_1, \ldots, a_K) \in H$. Then, for every $i \in N$, $s_i \in S_i(h) \equiv \text{proj}_{S_i} S(h)$ if and only if, for every $k = 1, \ldots, K$, if $P\left((a_1, \ldots, a_{k-1})\right) = i$ and $I \in \mathscr{I}_i$ is the unique information set such that $(a_1, \ldots, a_{k-1}) \in I$, then $s_i(I) = a_k$. In particular, $S(h) = \prod_{i \in N} S_i(h)$.*

[17] Osborne and Rubinstein (1994) start with a set H of histories, explicitly assume that this set is closed under the "sub-history" (prefix) relation, and define Z as the set of histories that are no proper prefix of any other history. I start from the set of terminal histories, and define H as the set of all prefixes (proper or not) thereof. This approach is more convenient in Definition 5, but equivalent.

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 $\mathscr{I}_j$ is a is a partition of $P^{-1}(\{j\}) \subseteq H \backslash Z$ for all $j \in N$, for every $k = 1, \ldots, K$, if $i = P\big((a_1, \ldots, a_{k-1})\big)$, then $(a_1, \ldots, a_{k-1}) \in I \in \mathscr{I}_j$ 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 $P\big((a_1, \ldots, a_{k-1})\big) = i$, $s_i(I) = a_k$, where $(a_1, \ldots, a_{k-1}) \in I \in \mathscr{I}_i$. Define $s_{-i} \in S_{-i}$ as follows: for every $j \neq i$ and all $J \in \mathscr{I}_j$, 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 $P\big((a_1, \ldots, a_{k-1})\big) = j$ and $(a_1, \ldots, a_{k-1}) \in J \in \mathscr{I}_j$ imply $s_j(J) = a_k$, regardless of whether $j = i$ or $j \neq i$. Hence, $(s_i, s_{-i}) \in S(h)$, so $s_i \in \text{proj}_{S_i} S(h) = S_i(h)$. ■

For completeness, I establish a (known) decomposition property of the sets $S(I)$, $I \in \mathscr{I}_i$, which also depends upon perfect recall. (I have been unable to find a published proof.)

Remark 3 *For all $i \in N$ and $I \in \mathscr{I}_i$, $S(I) = S_i(I) \times S_{-i}(I)$.*

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 2, this implies that $(s_i, s_{-i}) \in S(h'')$, and so $(s_i, s_{-i}) \in S(I)$, as claimed. ■

B Appendix: Proofs of the main results

B.1 Proof of Theorem 1

Suppose that $s_i \in S_i$ is structurally rational given μ. Consider an information set $I \in \mathscr{I}_i$ with $s_i \in S_i(I)$ and another strategy $r_i \in S_i(I)$. By strategic independence (cf. Sec. 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 Remark 1, there is a perturbation (p^k) of μ such that $U_i(s_i, p^k) \geq U_i(t_i, p^k)$. For this perturbation,

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

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.2 Proof of Theorem 2

Assume that $s_i \in S_i$ is weakly sequentially rational given μ, and that the game has no relevant ties for i. Fix an arbitrary $t_i \in S_i$.

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 $h \leq \zeta(t_i, s_{-i})$. If $h(s_{-i}) = \zeta(s_i, s_{-i})$, then also $h(s_i) = \zeta(t_i, s_{-i})$ and conversely, because terminal histories are not ranked by the prefix relation. 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 $J \in \mathscr{I}_j$ and $(h(s_{-i}), s_j(J)) \leq \zeta(s_i, s_{-i}), \zeta(t_i, s_{-i})$, which contradicts the definition of $h(s_{-i})$. Hence, either $h(s_{-i}) \in Z$, in which case $h(s_{-i}) = \zeta(s_i, s_{-i}) = \zeta(t_i, s_{-i})$, or else $h(s_{-i}) \in I$ for some $I \in \mathscr{I}_i$; in the latter case, denote the unique element of $\mathscr{I}_i$ 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 otherwise $a = s_i(I(s_{-i})) = t_i(I(s_{-i}))$ would satisfy $(h(s_{-i}, a) \leq \zeta(s_i, s_{-i})$ and $(h(s_{-i}), a) \leq \zeta(t_i, s_{-i})$, contradiction.

Consider $s_{-i}, t_{-i} \in S_{-i}$. I claim that $S_{-i}(I(s_{-i}))$ and $S_{-i}(I(t_{-i}))$ are either disjoint, or the same conditioning event. 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 $h'' \in I(s_{-i})$. Again by perfect recall, it must then be the case that $s_i(I(s_{-i})) = t_i(I(s_{-i}))$; however, as was shown above, $s_i(I(s_{-i})) \neq t_i(I(s_{-i}))$: contradiction. Similarly, it cannot be that $h' < h$. Thus, $h = h'$, and so $h = h' \in I(s_{-i}) \cap I(t_{-i})$. Since $\mathscr{I}_i$ partitions $P^{-1}(\{i\})$, $I(s_{-i}) = I(t_{-i})$. Therefore, writing $S^0_{-i} = \{s_{-i} : h(s_{-i}) \in Z\}\big\}$ and arbitrarily enumerating the collection $\{I(s_{-i}) : s_{-i} \in S_{-i}\}$ as $I_1, \ldots, I_L$, $\big\{S^0_{-i}\big\} \cup \big\{S_{-i}(I_\ell) : \ell = 1, \ldots, L\big\}$ is a partition of S_{-i}.

For all $s_{-i} \in S^0_{-i}$, by definition $U_i(s_i, s_{-i}) = U_i(t_i, s_{-i})$. By weak sequential rationality, for all $\ell = 1, \ldots, L$, $U_i\big(s_i, \mu(\cdot|I_\ell\big) \geq U_i\big(s_i, \mu(\cdot|I_\ell\big)$. Furthermore, fix one such ℓ. Since the game has no relevant ties and $s_i(I_\ell) \neq t_i(I_\ell)$, , for all $t_{-i} \in S_{-i}(I_\ell)$, either $U_i(s_i, t_{-i}) > U_i(t_i, t_{-i})$ or $U_i(s_i, t_{-i}) < U_i(t_i, t_{-i})$. Write $S^+_{-i}(I_\ell)$ and, respectively, $S^-_{-i}(I_\ell)$, for the collection of t_{-i} that satisfy the first or, respectively, the second inequality. Then

$$\sum_{t_{-i} \in S^+_{-i}(I_\ell)} \mu(\{t_{-i}\}|I_\ell)[U_i(s_i, t_{-i}) - U_i(t_i, t_{-i})] \geq \sum_{t_{-i} \in S^-_{-i}(I_\ell)} \mu(\{t_{-i}\}|I_\ell)[U_i(t_i, t_{-i}) - U_i(s_i, t_{-i})] \geq 0, \quad (13)$$

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

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

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

is well-defined. By Equation (13), $\lim_{k\to\infty}\alpha^k_\ell\leq 1$. Let $\beta^k_\ell=\max(\alpha^k(s_{-i}),1)$, so $\beta^k_\ell\geq 1$ and $\beta^k_\ell\to 1$; let $c=\left(p^k(S^0_{-i})+\sum_{m=1}^L[\beta^k_m p^k(S^+_{-i}(I_m))+p^k(S^-_{-i}(I_m))]\right)^{-1}$. Finally, define $(\tilde{p}^k)_{k\geq 1}$ by

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

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

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

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

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 $I \in \mathscr{I}_i$ 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 $\tilde{p}^k(\cdot|S_{-i}(I)) \to \mu(\cdot|I)$. Thus, by Remark 1, s_i is structurally rational given μ.

B.3 Elicitation

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

For $s^* \in S^*$, let $s^*_{-0i} = (s^*_j)_{j \in N \setminus \{i\}}$: that is, in addition to player i's strategy, we also disregard the experimenter's (player 0's) strategy. Similarly, let $S^*_{-0i} = \prod_{j \in N \setminus \{i\}} S^*_j$.

The following Lemma shows how to represent the set $S^*(I^*)$, for $I^* \in \cup_{i \in N} \mathscr{I}^*_i$, in terms of the functions $\bar{\mathbf{s}}_j(\cdot)$, $\mathbf{w}_i(\cdot)$, and $\mathbf{s}_i(\cdot)$.

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

$$S^*(I_{\bar{s}_i, w_i}) = \{i\} \times \left\{s^*_i : \bar{\mathbf{s}}_i(s^*_i) = \bar{s}_i, \mathbf{w}_i(s^*_i) = w_i, \mathbf{s}_i(s^*_i) \in S_i(I)\right\} \times \left\{s^*_{-0i} : \left(\bar{\mathbf{s}}_j(s^*_j)\right)_{j \in N \setminus \{i\}} \in S_{-i}(I)\right\}. \quad (14)$$

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

Claim: Let $h^* = \left(n, (\bar{t}_1, v_1), \ldots, (\bar{t}_i, v_i), \ldots, (\bar{t}_i, v_i), h\right) \in N \times \prod_{i \in N}(S_i \times W_i) \times (H \setminus Z)$. 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 $\big(n,(\bar{t}_1,v_1),\ldots,(\bar{t}_i,v_i),\ldots,(\bar{t}_i,v_i),z\big) \in Z^*$ 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 $z^* \equiv \big(i,(\bar{t}_1,v_1),\ldots,(\bar{t}_i,v_i),\ldots,(\bar{s}_i,w_i),z\big) \in Z^*$, 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,

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

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 $\big(\bar{\mathbf{s}}_j(s_j^*),\mathbf{w}_j(s_j^*)\big) = s_j^*(I_j^1) = (\bar{t}_j,v_j)$ for $j \in N \setminus \{i\}$, so $\big(\bar{\mathbf{s}}_j(s_j^*)\big)_{j\in N\setminus\{i\}} = \bar{t}_{-i} \in S_{-i}(h) \subseteq S_{-i}(I)$. Also, $\big(\bar{\mathbf{s}}_i(s_i^*),\mathbf{w}_i(s_i^*)\big) = s_i^*(I_i^1) = (\bar{s}_i,w_i)$.

In addition, let $h = (a_1,\ldots,a_K)$, and consider $k \in \{1,\ldots,K\}$ such that $P\big((a_1,\ldots,a_{k-1})\big) = i$. Let $J \in \mathscr{I}_i$ be such that $(a_1,\ldots,a_{k-1}) \in J$, and define $h_{k-1}^* = \big(i,(\bar{t}_1,v_1),\ldots,(\bar{s}_i,w_i),\ldots,(\bar{t}_N,v_N),a_1,\ldots,a_{k-1}\big)$. As noted above, there is t_i such that $(t_i,\bar{t}_{-i}) \in S(h) \subseteq S\big((a_1,\ldots,a_{k-1})\big)$. Therefore, by the Claim, $h_{k-1}^* \in J_{\bar{s}_i,w_i}$. Then, the definition of $\mathbf{s}_i(\cdot)$ and the fact that $s^* \in S^*(h^*)$ imply that $\mathbf{s}_i(s_i^*)(J) = s_i^*(J_{\bar{s}_i,w_i}) = a_k$. By Remark 2, $\mathbf{s}_i(s_i^*) \in S_i(h) \subseteq S_i(I)$. Therefore s^* belongs to the right-hand side of Eq. (14).

Conversely, suppose s^* belongs to the right-hand side of Eq. (14). By assumption $\big(\bar{\mathbf{s}}_j(s_j^*)\big)_{j\in N\setminus\{i\}} \in S_{-i}(I)$ and $\mathbf{s}_i(s_i^*) \in S_i(I)$, so by perfect recall $\big(\mathbf{s}_i(s_i^*),(\bar{\mathbf{s}}_j(s_j^*))\big) \in S(I)$. Hence there is $h \in I$ such that $\big(\mathbf{s}_i(s_i^*),(\bar{\mathbf{s}}_j(s_j^*))_{j\in N\setminus\{i\}}\big) \in S(h)$. Let

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

By assumption $s_0^*(\phi^*) = i$, $\bar{\mathbf{s}}_i(s_i^*) = \bar{s}_i$, and $\mathbf{w}_i(s_i^*) = w_i$. Furthermore, $\big(\mathbf{s}_i(s_i^*),(\bar{\mathbf{s}}_j(s_j^*))_{j\in N\setminus\{i\}}\big) \in S(h)$—that is, one can take $t_i = \mathbf{s}_i(s_i^*)$ 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 $h^* = (a_1^*, \ldots, a_K^*)$, with $K \geq N+1$. Thus, $h = (a_{N+2}^*, \ldots, a_K^*)$.[18] According to the definition, it must be shown that, for all $k = 1, \ldots, K$, action a_k^* is specified by s^* at history $(a_1^*, \ldots, a_{k-1}^*)$. 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 $(a_1^*, \ldots, a_{k-1}^*) = \left(s_0^*(\phi^*), (\bar{\mathbf{s}}_1(s_1^*), \mathbf{w}_1(s_1^*)), \ldots, (\bar{\mathbf{s}}_{k-2}(s_{k-2}^*), \mathbf{w}_{k-2}(s_{k-2}^*)) \in I_{k-1}^1\right.$ and, by the definition of h^*, $\bar{\mathbf{s}}_i(\cdot)$, and $\mathbf{w}_i(\cdot)$, $s_{k-1}^*(I_{k-1}^1) = (\bar{\mathbf{s}}_{k-1}(s_{k-1}^*), \mathbf{w}_{k-1}(s_{k-1}^*)) = a_k^*$.

If instead $k > N+1$, then $(a_1^*, \ldots, a_{k-1}^*) = \left(s_0^*(\phi^*), (\bar{\mathbf{s}}_1(s_1^*), \mathbf{w}_1(s_1^*)), \ldots, (\bar{\mathbf{s}}_N(s_N^*), \mathbf{w}_N(s_N^*)), a_{N+2}^*, \ldots, a_{k-1}^*\right)$, where $h' \equiv (a_{N+2}^*, \ldots, a_{k-1}^*) < (a_{N+2}^*, \ldots, a_k^*) \leq h$.[19] There are two sub-cases.

If $P(h') = i$, then also $P^*\left((a_1^*, \ldots, a_{k-1}^*)\right) = i$, and there exists $J \in \mathscr{I}_i$ such that $h' \in J$. Furthermore, $s_0^*(\phi^*) = i$, $\bar{\mathbf{s}}_i(s_i^*) = \bar{s}_i$, $\mathbf{w}_i(s_i^*) = w_i$, and $\left(\mathbf{s}_i(s_i^*), (\bar{\mathbf{s}}_j(s_j^*))_{j \in N \setminus \{i\}}\right) \in S(h) \subseteq S(h')$. Therefore, by the Claim, $(a_1^*, \ldots, a_{k-1}^*) \in J_{\bar{s}_i, w_i}$. Also, by Remark 2, $\mathbf{s}_i(s_i^*) \in S_i(h)$ implies $\mathbf{s}_i(s_i^*)(J) = a_k^*$. Conclude that $s_i^*(J_{\bar{s}_i, w_i}) = \mathbf{s}_i(s_i^*)(J) = a_k^*$.

If instead $P(h') = j \neq i$, then as above there is $J \in \mathscr{I}_j$ with $h' \in J$. In this case $\left(\mathbf{s}_i(s_i^*), (\bar{\mathbf{s}}_j(s_j^*))_{j \in N \setminus \{i\}}\right) \in S(h) \subseteq S(h')$ implies that $\bar{\mathbf{s}}_j(s_j^*)(J) = a_k^*$. Moreover, now $P^*\left((a_1^*, \ldots, a_{k-1}^*)\right) = 0$, and $(a_1^*, \ldots, a_{k-1}^*)$ is contained in the singleton information set $J^* = \{(a_1^*, \ldots, a_{k-1}^*)\} \in \mathscr{I}_0^*$. Now suppose that $a \in A$ is such that

$$(a_1^*, \ldots, a_{k-1}^*, a) = \left(s_0^*(\phi^*), (\bar{\mathbf{s}}_1(s_i^1), \mathbf{w}_1(s_i^*)), \ldots, (\bar{\mathbf{s}}_i(s_i^*), \mathbf{w}_i(s_i^*)), \ldots, (\bar{\mathbf{s}}_N(s_N^*), \mathbf{w}_N(s_N^*)), a_{N+2}^*, \ldots, a_{k-1}^*, a\right) \in H^*.$$

Then $(a_1^*, \ldots, a_{k-1}^*, a) < z^*$ for some $z^* \in Z^*$, and there must exist $z \in Z$ such that

$$z^* = \left(s_0^*(\phi^*), (\bar{\mathbf{s}}_1(s_i^1), \mathbf{w}_1(s_i^*)), \ldots, (\bar{\mathbf{s}}_i(s_i^*), \mathbf{w}_i(s_i^*)), \ldots, (\bar{\mathbf{s}}_N(s_N^*), \mathbf{w}_N(s_N^*)), z\right).$$

This requires that $(h', a) = (a_{N+2}^*, \ldots, a_{k-1}^*, a) < z$. In addition, the definition of Z^* requires that $(\bar{\mathbf{s}}_j(s_j^*))_{j \in N \setminus \{i\}} \in S_{-i}(z)$ (recall that $s_0^*(\phi^*) = i$), so by Remark 2, in particular $\bar{\mathbf{s}}_j(s_j^*)(J) = a$. But then $a = a_k^*$. Conclude that $A(J^*) = \{a_k^*\}$, so necessarily $s_0^*(J^*) = a_k^*$, as needed. ■

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

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

For every $s_{-i} \in S_{-i}$, let $[s_{-i}] = \{t^*_{-0i} \in S^*_{-0i} : \forall\, j \neq i, \bar{\mathbf{s}}_j(t^*_j) = s_j\}$. The collection $\{[s_{-i}] : s_{-i} \in S_{-i}\}$ partitions S^*_{-0i}. Furthermore, from Eq. (14),

$$S^*_{-i}(I_{\bar{s}_i, w_i}) = \{i\} \times \bigcup_{s_{-i} \in S_{-i}(I)} [s_{-i}]. \tag{15}$$

I now show that, for every CCPS in the original game, there is a CCPS in the elicitation game that agrees with it.

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 $I \in \mathscr{I}_i$ 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, $s^*_i(\bar{s}_i, w_i, s_i)$ reports the intended strategy $\bar{s}_i$ and bet w_i in the first stage, and then, if called upon to play directly in the second stage, 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 $\bar{\mathbf{s}}_i(s^*_i(\bar{s}_i, w_i, s_i) = \bar{s}_i$, $\mathbf{w}_i(s^*_i(\bar{s}_i, w_i, s_i) = w_i$, and $\mathbf{s}_i(s^*_i(\bar{s}_i, w_i, s_i) = s_i$.

Lemma 2 *For all $\mu_i \in \Delta(S_{-i}, \mathscr{I}_i)$ there is $\mu^*_i \in \Delta(S^*_{-i}, \mathscr{I}^*_i)$ 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 $\mu^*_i \in \Delta(S^*_{-i})^{\mathscr{I}^*_i}$ by letting, for every $n \in N$, $w_{-i} \in W_{-i}$, and $s_{-i} \in S_{-i}$,

$$\mu^*_i(\{s^*_{-i}(n, s_{-i}, w_{-i})\}|\phi^*) = \frac{1}{N \cdot |W_{-i}|}\mu(\{s_{-i}\}|\phi) \quad \text{and} \quad \forall\, I_{\bar{s}_i, w_i} \in \mathscr{I}^*_i,\ \mu^*_i(\{s^*_{-i}(i, s_{-i}, w_{-i})\}|I_{\bar{s}_i, w_i}) = \frac{1}{|W_{-i}|}\mu(\{s_{-i}\}|I),$$

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 $I^*_i \in \mathscr{I}^*_i$.

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 $I \in \mathscr{I}_i$ 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 Eq. (14) and the fact that $q^k(S^{**}_{-i}) = 1$,

$$q^k(\{s^*_{-i}(i,s_{-i},w_{-i})\}|S^*_{-i}(I_{\bar{s}_i,w_i})) = \frac{q^k(\{s^*_{-i}(i,s_{-i},w_{-i})\})}{\sum_{t_{-i}\in S_{-i}(I)} q^k(\{i\}\times[t_{-i}])} = \frac{q^k(\{s^*_{-i}(i,s_{-i},w_{-i})\})}{\sum_{t_{-i}\in S_{-i}(I),\tilde{w}_{-i}\in W_{-i}} q^k(s^*_{-i}(i,t_{-i},w_{-i}))}$$

$$= \frac{\frac{1}{N\cdot|W_{-i}|}p^k(\{s_{-i}\})}{\sum_{t_{-i}\in S_{-i}(I),\tilde{w}_{-i}\in W_{-i}} \frac{1}{N\cdot|W_{-i}|}p^k(\{t_{-i}\})} = \frac{1}{|W_{-i}|}p^k(\{s_{-i}\}|S_{-i}(I)) \to \frac{1}{|W_{-i}|}\mu_i(\{s_{-i}\}|I) = \mu^*_i(\{s^*_{-i}(n,s_{-i},w_{-i})\}|I_{\bar{s}_I,w_I}).$$

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

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

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

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

which completes the proof. ■

The following key lemma shows that one can obtain a perturbation of a CCPS μ_i in the original game from a perturbation of a CCPS μ^*_i in the elicitation game that agrees with it, and conversely. This is essential to relate structural preferences in the two games.

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

1. *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*

$$p^k(\{s_{-i}\}) = q^{k+\kappa-1}\left(\{i\}\times[s_{-i}]\middle|\{i\}\times S^*_{-0i}\right) \quad s_{-i}\in S_{-i}, k\geq 1 \tag{16}$$

is a perturbation of μ_i.

2. *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 Eq.* (16) *with* $\kappa = 1$*.*

Proof: For (1), by Eq. (11) 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 $I \in \mathscr{I}_i$. Then, for every $k \geq 1$, fixing an arbitrary $(\bar{s}_i, w_i) \in A(I_i^1) = S_i \times W_i$,

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

the last equality follows from Eq. (15), and the strict inequality from the assumption that $(q^k)_{k\geq 1}$ is a perturbation of μ_i^*. Furthermore, for every $s_{-i} \in S_{-i}(I)$, since by Eq. (15) $\{i\}\times[s_{-i}] \subseteq S_{-i}^*(I_{\bar{s}_i,w_i})$,

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

the third equality follows from the assumption that $(q^k)_{k\geq 1}$ is a perturbation of μ_i^*, and the last from agreement, i.e., Eq. (12) in Definition 6.

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

$$\lim_{k\to\infty} p^k(\{s_{-i}\}) = \lim_{k\to\infty} q^k(\{i\}\times[s_{-i}]|\{i\}\times S_{-0i}^*) = \lim_{k\to\infty} \frac{q^k(\{i\}\times[s_{-i}])}{q^k(\{i\}\times S_{-0i}^*)} = \frac{\lim_{k\to\infty} q^k(\{i\}\times[s_{-i}])}{\lim_{k\to\infty} q^k(\{i\}\times S_{-0i}^*)} =$$
$$= \frac{\mu_i^*(\{i\}\times[s_{-i}]|\phi^*)}{\mu_i^*(\{i\}\times S_{-0i}^*|\phi^*)} = \frac{\frac{1}{N}\mu_i(\{s_{-i}\}|\phi)}{\frac{1}{N}} = \mu_i(\{s_{-i}\}|I):$$

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 Eq. (11).

For (2), for every $I^* \in \mathscr{I}_i^* \cup \{\phi^*\}$, let

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

Fix $s^*_{-i} \in S^*_{-i}$ and let $s_{-i} = (\bar{\mathbf{s}}_j(s^*_j))_{j\in N\setminus\{i\}}$; 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 $I^*, J^* \in \mathscr{I}^*_i$. 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 Eq. (14), $s^*_{-i} \in \{s^*_0\}\times[s_{-i}] \subseteq S^*_{-i}(I^*)\cap S^*_{-i}(J^*)$.[20] 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 $s_{-i} = (\bar{\mathbf{s}}_j(s^*_j))_{j\in N\setminus\{i\}}$ and

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

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 $I^* \in \mathscr{I}^*_i$. 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

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

[20] This implies that, if e.g. $I^* = I_{s_i,w_i}$ 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. Similarly for J^*. However, this difference is immaterial to the argument in this paragraph.

i.e., Eq. (16) holds.

It remains to be shown that $(q^k)_{k\geq 1}$ is a perturbation of μ_i^*. For every $I^* \in \mathscr{I}_i^*$, 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 $I \in \mathscr{I}_i$. 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 Eq. (15), $q^k(S_{-i}^*(I_{\bar{s}_i,w_i})) \geq q^k(\{i\}\times[s_{-i}]) > 0$.

Now consider $I^* \in \{\phi^*, I_i^1\}$. Fix $s_{-i}^* \in S_{-i}^*$ and let $s_{-i} = (\bar{\mathbf{s}}_j(s_j^*))_{j\in N\setminus\{i\}}$. 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 $J^* \in \mathscr{I}_i^*$ 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 $I \in \mathscr{I}_i$ 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 $s_{-i} = (\bar{\mathbf{s}}_j(s_j^*))_{j\in N\setminus\{i\}}$ and $t_{-i} = (\bar{\mathbf{s}}_j(t_j^*))_{j\in N\setminus\{i\}}$. 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

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

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

$$\frac{q^k(\{s_{-i}^*\})}{q^k(\{t_{-i}^*\})}\leq\frac{q^k(\{i\}\times[s_{-i}])}{q^k(\{t_{-i}^*\})}\leq\frac{q^k(\{i\}\times[s_{-i}]|\{i\}\times S_{-0i}^*)}{q^k(\{t_{-i}^*\})}=\frac{p^k(\{s_{-i}\})}{p^k(\{t_{-i}\})\cdot\frac{1}{N}\cdot\rho(t_{-i}^*;I^*)}=$$
$$=\frac{N}{\rho(t_{-i}^*;I^*)}\cdot\frac{p^k(\{s_{-i}\}|S_{-i}(I))}{p^k(\{t_{-i}\}|S_{-i}(I))}\to\frac{N}{\rho(t_{-i}^*;I^*)}\cdot\frac{\mu_i(\{s_{-i}\}|I)}{\mu_i(\{t_{-i}\}|I)}=0=\frac{\mu_i^*(\{s_{-i}^*\}|I^*)}{\mu_i^*(\{t_{-i}^*\}|I^*)}.$$

The first equality is from Eq. (16); the limit statement follows because $(p^k)_{k\geq 1}$ is a perturbation of μ_i, and the last equality follows from $\mu_i^*(\{s_{-i}^*\}|I^*)\leq\mu_i^*(\{i\}\times[s_{-i}]|I^*)=0$.

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,

$$q^k(\{s_{-i}^*\}|S_{-i}^*(I^*))=\frac{q^k(\{s_{-i}^*\})}{\sum_{r_{-i}^*\in S_{-i}^*(I^*)}q^k(\{r_{-i}^*\})}=\frac{\frac{q^k(\{s_{-i}^*\})}{q^k(\{t_{-i}^*\})}}{\sum_{r_{-i}^*\in S_{-i}^*(I^*)}\frac{q^k(\{r_{-i}^*\})}{q^k(\{t_{-i}^*\})}}\to$$
$$\to\frac{\frac{\mu_i^*(\{s_{-i}^*\}|I^*)}{\mu_i^*(\{t_{-i}^*\}|I^*)}}{\sum_{r_{-i}^*\in S_{-i}^*(I^*)}\frac{\mu_i^*(\{r_{-i}^*\}|I^*)}{\mu_i^*(\{t_{-i}^*\}|I^*)}}=\frac{\mu_i^*(\{s_{-i}^*\}|I^*)}{\sum_{r_{-i}^*\in S_{-i}^*(I^*)}\mu_i^*(\{r_{-i}^*\}|I^*)}=\mu_i^*(\{s_{-i}^*\}|I^*),$$

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

Now rewrite the strategic-form payoff function in the elicitation game as follows. Fix $s^*\in S^*$, and let $z^*\in Z^*$ be such that $s^*\in S^*(z^*)$. By the definition of the maps $\bar{\mathbf{s}}_j(\cdot)$ and $\mathbf{w}_j(\cdot)$ for all $j\in N$, letting $n=s_0^*(\phi^*)$, $z^*=\left(n,(\bar{\mathbf{s}}_j(s_j^*),\mathbf{w}_j(s_j^*))_{j\in N},z\right)\in Z^*$, where $\bar{\mathbf{s}}_j(s_j^*)\in S_j(z)$ for all $j\in N\setminus\{n\}$. In addition, write $z=(a_1,\ldots,a_L)$, fix $K\in\{1,\ldots,L-1\}$, and let $h=(a_1,\ldots,a_K)$. Suppose that $P(h)=n$, so $h\in I\in\mathscr{I}_n$. Then $h^*\equiv\left(n,(\bar{\mathbf{s}}_j(s_j^*),\mathbf{w}_j(s_j^*))_{j\in N},h\right)\in H^*$, $P^*(h^*)=n$, 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 (10), $\mathbf{s}_n(s_n^*)(I)=s_n^*(I_{\bar{s}_n,w_n})=a_{K+1}$. Thus, for all K such that $P((a_1,\ldots,a_K))=n$, if $(a_1,\ldots,a_K)\in I\in\mathscr{I}_n$ then $\mathbf{s}_n(s_n^*)(I)=a_{K+1}$. By Remark 2, $\mathbf{s}_n(s_n^*)\in S_n(z)$, and so $\left(\mathbf{s}_n(s_n^*),(\bar{\mathbf{s}}_j(s_j^*))_{j\in N\setminus\{n\}}\right)\in S(z)$, i.e., $z=\zeta\left(\mathbf{s}_n(s_n^*),\left(\bar{\mathbf{s}}_j(s_j^*)\right)_{j\in N\setminus\{n\}}\right)$.

With this, for every $i \in N$, if $n \neq i$ then $U_i^*(s^*) = 0$; if $n = i$, since $u_i(z) = U_i\left(\mathbf{s}_i(s_i^*), \left(\bar{\mathbf{s}}_j(s_j^*)\right)_{j\in N\setminus\{i\}}\right)$,

$$U_i^*(s^*) = u_i^*(z^*) = \frac{1}{3}U_i\left(\mathbf{s}_i(s_i^*), \left(\bar{\mathbf{s}}_j(s_j^*)\right)_{j\in N\setminus\{i\}}\right) + \frac{1}{3}B_i\left(\mathbf{w}_i(s_i^*), \left(\bar{\mathbf{s}}_j(s_j^*)\right)_{j\in N\setminus\{i\}}\right) + \frac{1}{3}\epsilon \cdot 1_{\bar{\mathbf{s}}_i(s_i^*)\in S_i\left(\zeta\left(\mathbf{s}_i(s_i^*), \left(\bar{\mathbf{s}}_j(s_j^*)\right)_{j\in N\setminus\{i\}}\right)\right)}.$$

This emphasizes that i's payoff, if selected, depends on s_{-0i}^* only through $\left(\bar{\mathbf{s}}_j(s_j^*)\right)_{j\in N\setminus\{i\}}$, the profile of co-players' intended strategies. Now $\bar{\mathbf{s}}_i(s_i^*) \in S_i\left(\zeta\left(\mathbf{s}_i(s_i^*), \left(\bar{\mathbf{s}}_j(s_j^*)\right)_{j\in N\setminus\{i\}}\right)\right)$ if and only if $\zeta\left(\bar{\mathbf{s}}_i(s_i^*), \left(\bar{\mathbf{s}}_j(s_j^*)\right)_{j\in N\setminus\{i\}}\right) = \zeta\left(\mathbf{s}_i(s_i^*), \left(\bar{\mathbf{s}}_j(s_j^*)\right)_{j\in N\setminus\{i\}}\right)$. Therefore, for all $s_i^* \in S_i^*$ and $q \in \Delta(S_{-i}^*)$,

$$U_i^*(s_i^*, q) = \frac{1}{3}\sum_{s_{-i}\in S_{-i}} q(\{i\}\times[s_{-i}])U_i(\mathbf{s}_i(s_i^*), s_{-i}) + \frac{1}{3}\sum_{s_{-i}\in S_{-i}} q(\{i\}\times[s_{-i}])\cdot B_i(\mathbf{w}_i(s_i^*), s_{-i}) + \tag{17}$$

$$+\frac{1}{3}\epsilon \cdot \sum_{\substack{s_{-i}\in S_{-i}:\\ \zeta(\mathbf{s}_i(s_i^*), s_{-i}) = \zeta(\bar{\mathbf{s}}_i(s_i^*), s_{-i})}} q(\{i\}\times[s_{-i}]). \tag{18}$$

Proof of Theorem 3: 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 turn to (1) – (3).

(1) Since by assumption $\bar{\mathbf{s}}_i(s_i^*) = \mathbf{s}_i(s_i^*) \equiv s_i$, $\bar{\mathbf{s}}_i(t_i^*) = \mathbf{s}_i(t_i^*) \equiv t_i$, and $\mathbf{w}_i(s_i^*) = \mathbf{w}_i(t_i^*)$, for every $q \in \Delta(S_{-i}^*)$ the second and third terms in Eq. (17) for $U_i^*(s_i^*, q)$ and $U_i^*(t_i^*, q)$ have the same value, so

$$U_i^*(s_i^*, q) - U_i^*(t_i^*, q) = \frac{1}{3}\sum_{s_{-i}} q(\{i\}\times[s_{-i}])[U_i(s_i, s_{-i}) - U_i(t_i, s_{-i})]. \tag{19}$$

Suppose that $s_i^* \succ^{\mu_i^*} t_i^*$, and let $(p^k)_{k\geq 1}$ be a perturbation of μ_i. Since μ_i^* agrees with μ_i, by Lemma 3 part (2), there is a perturbation $(q^k)_{k\geq 1}$ of μ_i^* that satisfies Eq. (16) with $\kappa = 1$. By definition of structural preference, $U_i^*(s_i^*, q^k) - U_i^*(t_i^*, q^k) > 0$ eventually. By Eq. (16), since $q^k(\{i\}\times S_{-0i}^*) > 0$ eventually as $(q^k)_{k\geq 1}$ is a perturbation of μ_i^*, for all $r_i \in S_i$, eventually

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

Hence, by (19), $U_i(s_i, p^k) - U_i(t_i, p^k) > 0$ eventually. Since $(p^k)_{k\geq 1}$ was arbitrary, $s_i \succ^{\mu_i} t_i$.

Conversely, suppose that $s_i \succ^{\mu_i} t_i$ and let $(q^k)_{k\geq 1}$ be a perturbation of μ_i^*. Since μ_i^* agrees with μ_i, by Lemma 3 part (1), there is a perturbation $(p^k)_{k\geq 1}$ of μ_i that satisfies Eq. (16). For

this perturbation, $U_i(s_i,p^k)-U_i(t_i,p^k)>0$ eventually. By Eq. (16), for every $r_i\in S_i$, eventually

$$\sum_{s_{-i}} q^{k+\kappa-1}(\{i\}\times[s_{-i}])\cdot U_i(r_i,s_{-i}) = q^{k+\kappa-1}(\{i\}\times S^*_{-0i})\sum_{s_{-i}} q^{k+\kappa-1}(\{i\}\times[s_{-i}]|\{i\}\times S^*_{-0i})\cdot U_i(r_i,s_{-i}) =$$
$$= q^{k+\kappa-1}(\{i\}\times S^*_{-0i})U_i(r_i,p^k),$$

so by (19), $U_i^*(s_i^*,q^k)-U_i^*(t_i^*,q^k)>0$ eventually. Since $(q^k)_{k\geq 1}$ was arbitrary, $s_i^*\succ^{\mu_i^*} t_i^*$.

(2) Since by assumption $\bar{\mathbf{s}}_i(s_i^*)=\bar{\mathbf{s}}_i(t_i^*)$, $\mathbf{s}_i(s_i^*)=\mathbf{s}_i(t_i^*)$, $\mathbf{w}_i(s_i^*)=p$ and $\mathbf{w}_i(t_i^*)=E$, for every $q\in\Delta(S^*_{-i})$ the first and third terms in Eq. (17) for $U_i^*(s_i^*,q)$ and $U_i^*(t_i^*,q)$ are equal, and

$$\begin{aligned}
U_i^*(s_i^*,q)-U_i^*(t_i^*,q) &= \frac{1}{3}\sum_{s_{-i}\in S_{-i}} q(\{i\}\times[s_{-i}])\big[B_i(p,s_{-i})-B_i(E,s_{-i})\big] = \\
&= \frac{1}{3}\left[\sum_{s_{-i}\in E} q(\{i\}\times[s_{-i}])(p-1)+\sum_{s_{-i}\in S_{-i}(I_i)\setminus E} q(\{i\}\times[s_{-i}])p\right] = \\
&= \frac{1}{3}\left[p\sum_{s_{-i}\in S_{-i}(I_i)} q(\{i\}\times[s_{-i}])-\sum_{s_{-i}\in E} q(\{i\}\times[s_{-i}])\right] = \\
&= \frac{1}{3}\big[p\cdot q(S^*_{-i}(I_{s_i,w_i}))-q(\{i\}\times\cup\{[s_{-i}]:s_{-i}\in E\})\big], \qquad (20)
\end{aligned}$$

where the last equality follows from Eq. (14).

I prove the result for $p>\mu_i(E|I)$; the case $\mu_i(E|I)>p$ is analogous. Since μ_i^* agrees with μ_i, $p>\mu_i^*(\{i\}\times\cup\{[s_{-i}]:s_{-i}\in E\}|I_{\bar{s}_i,p})=\mu_i^*(\{i\}\times\cup\{[s_{-i}]:s_{-i}\in E\}|I_{\bar{s}_i,E})$ as well. Therefore, for any perturbation $\{q^k\}_{k\geq 1}$ of μ_i^*, and all $w_i\in W_i$,

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

the last equality uses the fact that, by Eq. (15), $E\subseteq S_{-i}(I)$ implies $\{i\}\times\cup\{[s_{-i}]:s_{-i}\in E\}\subseteq S^*_{-i}(I_{\bar{s}_i,w_i})$. Hence, for large k, $p\cdot q^k(S^*_{-i}(I_{\bar{s}_i,w_i}))-q^k(\{i\}\times\cup\{[s_{-i}]:s_{-i}\in E\})>0$; hence, by Eq. (20), $U_i^*(s_i^*,q^k)>U_i^*(t_i^*,q^k)$. Since (q^k) was an arbitrary perturbation of μ_i^*, $s_i^*\succ^{\mu_i^*} t_i^*$, as claimed.

(3) Let $z\in Z$ be such that $\bar{s}_i\equiv\bar{\mathbf{s}}_i(s_i^*)\notin S_i(z)$ and $s_i\equiv\mathbf{s}_i(s_i^*)=\mathbf{s}_i(t_i^*)=\bar{\mathbf{s}}_i(t_i^*)\in S_i(z)$. Since also $\mathbf{w}_i(s_i^*)=\mathbf{w}_i(t_i^*)$, for all $q\in\Delta(S^*_{-i})$ the first and second terms in Eq. (17) for $U_i^*(s_i^*,q)$ and

$U_i^*(t_i^*,q)$ are the same, and so

$$U_i^*(s_i^*,q)-U_i^*(t_i^*,q)=\frac{1}{3}\epsilon\left(\sum_{\substack{s_{-i}\\ \zeta(\mathbf{s}_i(s_i^*),s_{-i})=\zeta(\bar{\mathbf{s}}_i(s_i^*),s_{-i})}} q(\{i\}\times[s_{-i}])-\sum_{\substack{s_{-i}\\ \zeta(\mathbf{s}_i(t_i^*),s_{-i})=\zeta(\bar{\mathbf{s}}_i(t_i^*),s_{-i})}} q(\{i\}\times[s_{-i}])\right)=$$
$$=\frac{1}{3}\epsilon\left(\sum_{\substack{s_{-i}:\\ \zeta(s_i,s_{-i})=\zeta(\bar{s}_i,s_{-i})}} q(\{i\}\times[s_{-i}])-\sum_{\substack{s_{-i}\\ \zeta(s_i,s_{-i})=\zeta(s_i,s_{-i})}} q(\{i\}\times[s_{-i}])\right)=$$
$$=\frac{1}{3}\epsilon\left(\sum_{\substack{s_{-i}:\\ \zeta(s_i,s_{-i})=\zeta(\bar{s}_i,s_{-i})}} q(\{i\}\times[s_{-i}])-1\right).$$

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 $\bar{s}_i\in S_i(z)$, contradiction. Let h be the last common prefix of z and $\zeta(\bar{s}_i,t_{-i})$, i.e., the longest non-terminal history such that $h<z$ and $h<\zeta(\bar{s},t_{-i})$. Let $h\in I\in\mathscr{I}_i$. 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

$$U_i^*(s_i^*,q)-U_i^*(t_i^*,q)=\frac{1}{3}\epsilon\left(\sum_{\substack{s_{-i}\\ \zeta(s_i,s_{-i})=\zeta(\bar{s}_i,s_{-i})}} q(\{i\}\times[s_{-i}])-1\right)\leq-\frac{1}{3}\epsilon\sum_{s_{-i}\in S_{-i}(I)} q(\{i\}\times[s_{-i}]).$$

Finally, for any $\nu_i^*\in\Delta(S_{-i}^*,\mathscr{I}_i^*)$ and any perturbation $(q^k)_{k\geq1}$ of ν_i^*, by Eq. (14),

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

Therefore, for all k, $U_i^*(t_i^*,q^k)>U_i^*(s_i^*,q^k)$. It follows that $t_i^*\succ^{\nu_i^*}s_i^*$, as claimed.

For the final claim, suppose that s_i^* is structurally rational given μ_i^*. Let $\bar{s}_i=\bar{\mathbf{s}}_i(s_i^*)$, $w_i=\mathbf{w}_i(s_i^*)$, and $s_i=\mathbf{s}_i(s_i^*)$.

I first prove the last part of the final claim: for every $z\in Z$, $s_i\in S_i(z)$ iff $\bar{s}_i\in S_i(z)$—that is, s_i and $\bar{s}_i$ are *realization-equivalent*. By (3), if $\bar{s}_i\notin S_i(z)$ but $s_i\in S_i(z)$, then (3) immediately implies that there is t_i^* such that, in particular $t_i^*\succ^{\mu_i^*}s_i^*$; thus, s_i^* cannot be structurally rational, contradiction. Suppose instead $\bar{s}_i\in S_i(z)$ but $s_i\notin S_i(z)$. Let $s_{-i}\in S_{-i}(z)$, so $(\bar{s}_i,s_{-i})\in S(z)$. It cannot be that $(s_i,s_{-i})\in S_i(z)$, for otherwise $s_i\in S_i(z)$. Thus, $z'\equiv\zeta(s_i,s_{-i})\neq z$. Then $s_i\in S_i(z')$

and $s_{-i} \in S_{-i}(z')$. Suppose also $\bar{s}_i \in S_i(z')$. Then, by Remark 2, since $s_{-i} \in S_{-i}(z')$, $(\bar{s}_i, s_{-i}) \in S(z')$, i.e., $z' = \zeta(\bar{s}_i, s_{-i}) = z$, contradiction. Hence, $\bar{s}_i \notin S_i(z')$, and again we conclude that s_i^* cannot be structurally rational, contradiction. Thus, s_i and $\bar{s}_i$ are realization-equivalent.

Consequently, for every $s_{-i} \in S_{-i}(z)$, by Remark 2, $(s_i, s_{-i}) \in S(z)$ iff $(\bar{s}_i, s_{-i}) \in S(z)$, so that $U_i(s_i, s_{-i}) = U_i(\bar{s}_i, s_{-i})$, and $U_i(s_i, p) = U_i(\bar{s}_i, p)$ for every $p \in \Delta(S_{-i})$. Hence, by the definition of structural preference, for all $t_i \in S_i$, $t_i \succ^{\mu_i} s_i$ iff $t_i \succ^{\mu_i} \bar{s}_i$. In particular, s_i is structurally rational given μ_i, if and only if $\bar{s}_i$ is.

In addition, by Eq. (17), this implies that, letting $\tilde{s}_i^* = s_i^*(s_i, w_i, s_i)$, for every $q \in \Delta(S_{-i}^*)$, $U_i^*(\tilde{s}_i^*, q) = U_i^*(s_i^*, q)$. Hence, by the definition of structural preferences, for all $t_i^* \in S_i^*$, $t_i^* \succ^{\mu_i^*} s_i^*$ iff $t_i^* \succ^{\mu_i^*} \tilde{s}_i^*$. In particular, since s_i^* is structurally rational given μ_i^*, so is $\tilde{s}_i^*$.

Now suppose s_i is not structurally rational given μ_i, so there is $t_i \in S_i$ with $t_i \succ^{\mu_i} s_i$. Let $t_i^* = s_i^*(t_i, w_i, t_i)$. Then $\mathbf{s}_i(t_i^*) = t_i \succ^{\mu_i} s_i = \mathbf{s}_i(\tilde{s}_i^*)$, so by (1), $t_i^* \succ^{\mu_i^*} \tilde{s}_i^*$. But this contradicts the fact that, as was just shown, $\tilde{s}_i^*$ is (also) structurally rational given μ_i^*. Therefore, s_i is structurally rational given μ_i, and hence so is $\bar{s}_i$.

Finally, suppose that $w_i = p$, so that $\tilde{s}_i^* = s_i^*(s_i, E, s_i)$. Suppose that $\mu_i(E|I_i) > p$, and let $t_i^* = s_i^*(s_i, E, s_i)$. Then, by (2), $t_i^* \succ^{\mu_i^*} \tilde{s}_i^*$, which contradicts the fact that $\tilde{s}_i^*$ is (also) structurally rational. The case $w_i = E$ is analogous, hence omitted. ■

References

R.J. Aumann and J.H. Dreze. Assessing strategic risk. *American Economic Journal: Microeconomics*, 1(1):1–16, 2009.

P. Battigalli. On rationalizability in extensive games. *Journal of Economic Theory*, 74(1):40–61, 1997.

P. Battigalli and M. Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106(2):356–391, 2002.

G.M. Becker, M.H. DeGroot, and J. Marschak. Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3), 1964.

E. Ben-Porath. Rationality, Nash equilibrium and backwards induction in perfect-information games. *The Review of Economic Studies*, pages 23–46, 1997.

Truman Bewley. Knightian decision theory: Part I. *Decisions in Economics and Finance*, 25(2): 79–110, November 2002. (first version 1986).

Mariana Blanco, Dirk Engelmann, Alexander K Koch, and Hans-Theo Normann. Belief elicitation in experiments: is there a hedging problem? *Experimental Economics*, 13(4):412–438, 2010.

Miguel A Costa-Gomes and Georg Weizsäcker. Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3):729–762, 2008.

Daniel Ellsberg. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75: 643–669, 1961.

Itzhak Gilboa and David Schmeidler. A derivation of expected utility maximization in the context of a game. *Games and Economic Behavior*, 44(1):172–182, 2003.

S. Govindan and R. Wilson. On forward induction. *Econometrica*, 77(1):1–28, 2009. ISSN 1468-0262.

Elon Kohlberg and Philip J Reny. Independence on relative probability spaces and consistent assessments in game trees. *Journal of Economic Theory*, 75(2):280–313, 1997.

D.M. Kreps and R. Wilson. Sequential equilibria. *Econometrica: Journal of the Econometric Society*, 50(4):863–894, 1982.

R. Duncan Luce and Howard Raiffa. *Games and Decisions*. Wiley, New York, 1957.

George J Mailath, Larry Samuelson, and Jeroen M Swinkels. Extensive form reasoning in normal form games. *Econometrica*, 61:273–302, 1993.

R.B. Myerson. Multistage games with communication. *Econometrica*, 54(2):323–358, 1986. ISSN 0012-9682.

Yaw Nyarko and Andrew Schotter. An experimental study of belief learning using elicited beliefs. *Econometrica*, 70(3):971–1005, 2002.

Martin J. Osborne and A. Rubinstein. *A Course on Game Theory*. MIT Press, Cambridge, MA, 1994.

P.J. Reny. Backward induction, normal form perfection and explicable equilibria. *Econometrica*, 60(3):627–649, 1992. ISSN 0012-9682.

A. Rényi. On a new axiomatic theory of probability. *Acta Mathematica Hungarica*, 6(3):285–335, 1955. ISSN 0236-5294.

Pedro Rey-Biel. Equilibrium play and best response to (stated) beliefs in normal form games. *Games and Economic Behavior*, 65(2):572–585, 2009.

A. Rubinstein. Comments on the interpretation of game theory. *Econometrica*, 59(4):909–924, 1991. ISSN 0012-9682.

Leonard J. Savage. *The Foundations of Statistics*. Wiley, New York, 1954.

R. Selten. Ein oligopolexperiment mit preisvariation und investition. *Beiträge zur experimentellen Wirtschaftsforschung, ed. by H. Sauermann, JCB Mohr (Paul Siebeck), Tübingen*, pages 103–135, 1967.

R. Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International journal of game theory*, 4(1):25–55, 1975. ISSN 0020-7276.

Marciano Siniscalchi. Foundations for sequential preferences. mimeo, Northwestern University, 2020a.

Marciano Siniscalchi. Structural rationality in dynamic games. mimeo, Northwestern University, May 2020b.

Marciano Siniscalchi. Putting structural rationality to work. mimeo, Northwestern University, October 2021.

Eric Van Damme. Stable equilibria and forward induction. *journal of Economic Theory*, 48(2): 476–496, 1989.