

## Refinements of rationalizability for normal-form games\*

P. Jean-Jacques Herings<sup>1</sup>, Vincent J. Vannetelbosch<sup>2,3</sup>

<sup>1</sup> CentER and Department of Econometrics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands (e-mail: P.J.J.Herings@kub.nl)

<sup>2</sup> CORE, University of Louvain, voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium

<sup>3</sup> IEP, Basque Country University, Avda. Lehendakari Aguirre 83, E-48015 Bilbao, Spain (e-mail: vv@bl.ehu.es)

Received: January 1997/final version: August 1998

**Abstract.** There exist three equivalent definitions of perfect Nash equilibria which differ in the way “best responses against small perturbations” are defined. It is shown that applying the spirit of these definitions to rationalizability leads to three different refinements of rationalizable strategies which are termed perfect (Bernheim, 1984), weakly perfect and trembling-hand perfect rationalizability, respectively. We prove that weakly perfect rationalizability is weaker than both perfect and proper (Schuhmacher, 1995) rationalizability and in two-player games it is weaker than trembling-hand perfect rationalizability. By means of examples, it is shown that no other relationships can be found.

**Key words:** Rationalizability, refinements

---

### 1. Introduction

A notion like Nash equilibrium assumes common expectations of the players' behaviour. That is, each player holds a correct conjecture about her opponents' strategy choice. But once we admit the possibility that a player may have several strategies that she could reasonably use, conjectures and strategies actually played may be mismatched. This is what distinguishes ration-

---

\* We would like to thank Eric van Damme, Claude d'Aspremont, Pierre Dehez, Françoise Forges, Hans Peters and three anonymous referees for helpful comments and suggestions. This paper has been presented at seminars or conferences at the University of Aarhus, Basque Country University, SUNY at Stony Brook, ASSET 97 Meeting in Marseille. The research of P.J.J. Herings has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

	$Y_1$	$Y_2$
$X_1$	5, 1	1, 5
$X_2$	3, 3	3, 3

**Fig. 1.** A two-player game: G1

alizability [Bernheim (1984), Pearce (1984)] from equilibrium concepts. But rationalizability for normal-form games on its own fails to exclude some implausible strategy choices. One example is the game given in Figure 1. It can be shown that  $\{(X_1, Y_1), (X_1, Y_2), (X_2, Y_1), (X_2, Y_2)\}$  are all rationalizable; in other words, all pure strategies are possible best responses and rationalizable in G1. However, action  $Y_1$  of player 2 is weakly dominated, and it seems natural to assume that players do not consider such inadmissible strategies. Therefore, one would like to have a solution concept that yields  $(X_2, Y_2)$  as the only solution of the game.

The notion of Nash equilibrium faces similar problems. To avoid unreasonable outcomes, many refinements of Nash equilibrium have been proposed in the literature. The basic idea behind refinements is that players make a mistake with a small probability. For perfect Nash equilibria equivalent definitions are obtained by either modelling these mistakes by the requirement that each pure strategy is chosen with some minimum probability, or by assuming that rational players make a mistake only with at most some small probability. The definition of perfect rationalizability given by Bernheim uses the first approach to get a refinement of rationalizability. We propose a new refinement, called weakly perfect rationalizability, by taking the second approach. This also closes the gap to proper rationalizability (Schuhmacher, 1995), where necessarily a definition using the second approach is taken.

Related ideas are used in the definition of cautious rationalizability (Pearce, 1984). A strategy of a player is said to be a cautious response if it is a best response against a completely mixed strategy combination. Cautiously rationalizable strategy combinations are obtained by eliminating strategies that are not best responses first, next those that are not cautious responses, then the ones that are not best responses, and so on. Cautious rationalizability seems to be in between rationalizability and perfect rationalizability. If one carries the logic behind cautious rationalizability one step further, one would like to consider a solution concept where players eliminate responses that are not cautious in each round, which leads us to the concept of trembling-hand perfect rationalizability. In this concept players' actions have to be best responses also against perturbed conjectures. It is also closely related to yet another definition of perfect Nash equilibrium that defines a perfect Nash equilibrium as the limit point of a sequence of completely mixed strategy combinations and being a best reply against every element in this sequence.

Based on the intuition derived from the equilibrium approach, the reader might expect perfect rationalizability to be equivalent to weakly perfect rationalizability and to trembling-hand perfect rationalizability. Furthermore, one may expect these concepts to be a coarsening of proper rationalizability and a refinement of cautious rationalizability.

The main results of the paper show this intuition to be wrong. Weakly perfect rationalizability is a weaker refinement than both perfect and proper rationalizability. Moreover, in two-player games it holds that weakly perfect rationalizability is a weaker refinement than trembling-hand perfect rationalizability. This result is not true for games with more than two players as will be shown by an example. For the relationship between any other two refinements we give examples showing that the remaining set of strategies corresponding to the first refinement can be either smaller or larger than the one corresponding to the second refinement.

Contrary to equilibrium concepts, the cutting power of refinements of rationalizability depends very much and sometimes in unexpected ways on how exactly mistakes and cautiousness are modelled. Trembling-hand rationalizability is the only refinement that is not vulnerable, in the sense of giving different solutions, to adding strictly dominated strategies to a game. Moreover, in many interesting examples like the burning-money game for instance, trembling-hand perfect rationalizability has most cutting power of all refinements.

The paper is organized as follows. In Section 2 the rationalizability concept and the existing refinements, i.e. perfect rationalizability, proper rationalizability, and cautious rationalizability, are presented. In Section 3 we give two new refinements, weakly perfect rationalizability and trembling-hand perfect rationalizability, which are obtained by applying the spirit of equivalent definitions of perfect Nash equilibrium to rationalizability. We derive the earlier mentioned relationships between the refinements in Section 4, and we show by means of examples in Section 5 that there are no other relationships.

## 2. Rationalizability and existing refinements

We consider a normal-form game  $\Gamma(I, S, U)$ , where  $I$  is a finite set of players. Each player  $i \in I$  has a finite pure-strategy set  $S_i$  and a payoff function  $U_i$ . We denote  $S = \prod_{i \in I} S_i$  and  $U = (U_i)_{i \in I}$ . Let  $M_i$  be the set of player  $i$ 's mixed strategies and  $M = \prod_{i \in I} M_i$  the set of mixed strategy combinations. Given  $c_i \in M_i$ , we denote by  $c_i(s_i)$  the probability that  $c_i$  assigns to pure strategy  $s_i$ . Player  $i$ 's opponents in the game  $\Gamma(I, S, U)$  are denoted by  $-i$ . The notation  $-i$  is also used to denote products over all players except  $i$ , for instance in  $c_{-i}$  or  $M_{-i}$ . As general notation, given any set  $X$ , we denote by  $\text{ch}(X)$  the convex hull of the set  $X$ , i.e. the smallest convex set containing  $X$ . For a subset  $X$  of a Euclidean space, we denote by  $\text{int}(X)$  the relative interior of the set  $X$ .

Rationalizability [Bernheim (1984), Pearce (1984)] for normal-form games is based on the following assumptions: **(A1)** the players are rational, **(A2)** **A1** is common knowledge among the players, and **(A3)** the structure of the game (strategy sets, payoff functions) is common knowledge. Our formulation of rationality is based on expected utility maximization given uncorrelated<sup>1</sup> conjectures about the opponents' strategies. Rationalizability for normal-form games can be defined by the following iterative process.

<sup>1</sup> Correlated rationalizability, introduced by Brandenburger and Dekel (1987), weakens rationalizability because allowing correlated conjectures about the strategies of the opponents makes more strategies rationalizable. In the paper, we only consider the case where the players hold uncorrelated conjectures.

	$Y_1$	$Y_2$
$X_1$	1, 1	0, 0
$X_2$	0, 0	0, 0

Fig. 2. A two-player game: G2

**Definition 1.** Let  $R^0 = M$ . For  $k \geq 1$ ,  $R^k = \prod_{i \in I} R_i^k$  is inductively defined as follows:  $c_i$  belongs to  $R_i^k$  if  $c_i \in M_i$  and there is a  $c_{-i} \in \text{ch}(R_{-i}^{k-1})$  such that  $c_i$  is a best response against  $c_{-i}$  within  $M_i$ . The set  $R^\infty = \lim_{k \rightarrow \infty} R^k$  is the set of rationalizable strategy profiles.

Consider the two-player normal-form game G2 (see Figure 2) from Myerson (1978). This game possesses two pure Nash equilibria:  $\{(X_1, Y_1), (X_2, Y_2)\}$ . Then, it is straightforward that  $\{(X_1, Y_1), (X_1, Y_2), (X_2, Y_1), (X_2, Y_2)\} \subset R^\infty$ , so all pure strategy profiles are rationalizable. Nonetheless, the pure strategy profiles  $(X_1, Y_2)$ ,  $(X_2, Y_1)$  and  $(X_2, Y_2)$  seem unreasonable, because they involve weakly dominated strategies.

To exclude these unreasonable outcomes, three refinements have been introduced in the literature: perfect rationalizability, proper rationalizability, and cautious rationalizability.

Perfect rationalizability is due to Bernheim (1984). The idea behind the perfectness notion is that each player makes mistakes with a small probability, which has the consequence that every pure strategy is chosen with a positive probability. It is assumed that these minimum probabilities are common knowledge. Strategies are perfectly rationalizable if they are the limit of rationalizable strategies in these perturbed games as the minimum probabilities in the perturbed games converge to zero.

Given a strictly positive vector  $\mu = (\mu_i)_{i \in I}$ , we denote by  $M_i(\mu)$  the set of strategies of player  $i$  that assign probabilities of at least  $\mu_i(s_i) > 0$  to pure strategies  $s_i$  of player  $i$ , so  $M_i(\mu) = \{c_i \in M_i \mid c_i(s_i) \geq \mu_i(s_i), \forall s_i \in S_i\}$ . Perfect rationalizability for normal-form games is defined by the following iterative procedure.

**Definition 2.** Let a strictly positive vector  $\mu$  be given and let  $B^0(\mu) = \prod_{i \in I} M_i(\mu)$ . For  $k \geq 1$ ,  $B^k(\mu) = \prod_{i \in I} B_i^k(\mu)$  is inductively defined as follows:  $c_i$  belongs to  $B_i^k(\mu)$  if  $c_i \in M_i(\mu)$  and there is a  $c_{-i} \in \text{ch}(B_{-i}^{k-1}(\mu))$  such that  $c_i$  is a best response against  $c_{-i}$  within  $M_i(\mu)$ . The set  $B^\infty(\mu) = \lim_{k \rightarrow \infty} B^k(\mu)$  is the set of  $\mu$ -perfectly rationalizable strategy profiles and  $B^\infty = \lim_{\mu \rightarrow 0^+} B^\infty(\mu)$  the set of perfectly rationalizable strategy profiles.

In Definition 2, the set  $B^\infty$  is given by

$$\lim_{\mu \rightarrow 0^+} B^\infty(\mu) = \{c \in M \mid \exists \{\mu^t\}_{t=0}^\infty \rightarrow 0^+, \exists \{c^t\}_{t=0}^\infty \rightarrow c, c^t \in B^\infty(\mu^t)\}.$$

Reconsider the two-player normal-form game G2 (see Figure 2). Recall that the pure strategy profiles  $(X_1, Y_2)$ ,  $(X_2, Y_1)$  and  $(X_2, Y_2)$  are rationalizable. Nevertheless, none of these pure strategy profiles are perfectly

rationalizable. Take any strictly positive  $\mu$ . It is obvious that  $B^1(\mu) = B^k(\mu)$  for all  $k > 1$  and  $c \in B^1(\mu)$  implies  $c_1(X_2) = \mu_1(X_2)$  and  $c_2(Y_2) = \mu_2(Y_2)$ . It follows that  $(X_1, Y_1)$  is the unique perfectly rationalizable strategy profile. As a more general property, one can verify that weakly dominated strategies are always eliminated by perfect rationalizability.

Schuhmacher (1995) has developed the proper rationalizability concept<sup>2</sup>. Proper rationalizability for normal-form games is defined by the following iterative procedure.

**Definition 3.** Let  $\varepsilon > 0$  be given and let  $A^0(\varepsilon) = \prod_{i \in I} \text{int}(M_i)$ . For  $k \geq 1$ ,  $A^k(\varepsilon) = \prod_{i \in I} A_i^k(\varepsilon)$  is inductively defined as follows:  $c_i \in A_i^k(\varepsilon)$  if  $c_i$  belongs to  $\text{int}(M_i)$  and there is a  $c_{-i} \in \text{ch}(A_{-i}^{k-1}(\varepsilon))$  such that for every two pure strategies  $s_i, s'_i \in S_i$  with  $s_i$  a strictly better response against  $c_{-i}$  than  $s'_i$ , it holds that  $c_i(s'_i) \leq \varepsilon c_i(s_i)$ . The set  $A^\infty(\varepsilon) = \lim_{k \rightarrow \infty} A^k(\varepsilon)$  is the set of  $\varepsilon$ -properly rationalizable strategy profiles and  $A^\infty = \lim_{\varepsilon \rightarrow 0^+} A^\infty(\varepsilon)$  the set of properly rationalizable strategy profiles.

As for perfect rationalizability, the reader can verify that weakly dominated strategies are eliminated by proper rationalizability. It is possible to give examples where the concept of proper rationalizability has more cutting power than perfect rationalizability. Consider example G3, taken from Myerson (1978), which highlights how the perfectness notion may fail to eliminate all intuitively unreasonable outcomes. There are three Nash equilibria, and all are in pure strategies; these equilibria are  $(X_1, Y_1)$ ,  $(X_2, Y_2)$ , and  $(X_3, Y_3)$ . Of these three Nash equilibria,  $(X_3, Y_3)$  is not perfect nor proper,  $(X_2, Y_2)$  is perfect but not proper, and  $(X_1, Y_1)$  is both perfect and proper. Theorem 1 states that perfect Nash equilibria are perfectly rationalizable strategy profiles, so  $(X_1, Y_1)$  and  $(X_2, Y_2)$  are perfectly rationalizable. We claim that  $(X_1, Y_1)$  is the unique properly rationalizable strategy profile of G3. Consider any  $\varepsilon \in (0, \frac{1}{2})$ . The reader may verify that

$$A^1(\varepsilon) = \{(c_1, c_2) \in \text{int}(M) \mid c_1(X_3) \leq \varepsilon c_1(X_2) \quad \text{and} \quad c_2(Y_3) \leq \varepsilon c_2(Y_2)\}.$$

In addition, for  $c$  to be a member of  $A^2(\varepsilon)$  it should hold that  $c_1(X_3) \leq \varepsilon c_1(X_1)$  and  $c_2(Y_3) \leq \varepsilon c_2(Y_1)$ . And for  $c$  to be a member of  $A^3(\varepsilon)$  it is required on top of this that  $c_1(X_2) \leq \varepsilon c_1(X_1)$  and  $c_2(Y_2) \leq \varepsilon c_2(Y_1)$ . Therefore, we have that  $A^\infty = \{(X_1, Y_1)\}$ . In the sections on general relationships between refinements, it will be shown that it is not always the case that  $A^\infty \subseteq B^\infty$ .

Cautious rationalizability, due to Pearce (1984), imposes the condition that the players' conjectures give positive weight to all rationalizable alternatives, whereas the strategy profiles that are not rationalizable should be given zero weight. Formally, cautious rationalizability is defined by the following iterative procedure.

**Definition 4.** Let  $C^0 = M$ . For  $k \geq 1$ ,  $C^k = \prod_{i \in I} C_i^k$  is inductively defined as follows:  $c_i \in C_i^k$  if  $c_i \in R_i^\infty(C^{k-1})$  and there is a  $c_{-i} \in \text{int}(\text{ch}(R_{-i}^\infty(C^{k-1})))$

<sup>2</sup> The properness notion has been first introduced by Myerson (1978), in the equilibrium approach, to refine the perfect equilibrium concept due to Selten (1975). Schuhmacher (1995) has shown that proper rationalizability implies the backward induction outcome for generic extensive-form games with perfect information.

	$Y_1$	$Y_2$	$Y_3$
$X_1$	1, 1	0, 0	-9, -9
$X_2$	0, 0	0, 0	-7, -7
$X_3$	-9, -9	-7, -7	-7, -7

Fig. 3. A two-player game: G3

such that  $c_i$  is a best response against  $c_{-i}$  within  $R_i^\infty(C^{k-1})$ . The set  $C^\infty = \lim_{k \rightarrow \infty} C^k$  is the set of cautiously rationalizable strategy profiles.

In Definition 4,  $R_i^\infty(C^{k-1})$  is player  $i$ 's set of rationalizable strategies given that the set of players' strategy profiles is  $C^{k-1}$ . At each step of the iterative procedure, strategies that are not best responses are eliminated first, and those that are not cautious responses, i.e. best responses against a completely mixed strategy profile, are removed next.

Let  $C$  be a solution concept, i.e.  $C$  is a function that assigns to each game a set of solutions of that game. An important property a solution concept may or may not satisfy is the pure strategy property.

**Definition 5.** Let  $C$  be a solution concept such that  $C(\Gamma) = \prod_{i \in I} C_i(\Gamma)$  for every game  $\Gamma$  with  $C_i(\Gamma) \subseteq M_i$ . The solution concept  $C$  has the pure strategy property if  $c_i(s_i) > 0$  for some  $c_i \in C_i(\Gamma)$  implies  $s_i \in C_i(\Gamma)$ .

For rationalizability and its refinements the pure strategy property is very important. It implies that it is sufficient to know the pure strategy combinations that are assigned as a solution to a game in order to determine all conjectures that a player can hold.

It is not difficult to show the following result, see also Bernheim (1984), Pearce (1984) and Schuhmacher (1995).

**Theorem 1.** For every normal-form game  $\Gamma(I, S, U)$ :

1. The sets of rationalizable, perfectly rationalizable, properly rationalizable and cautiously rationalizable strategy profiles are non-empty and closed. All of these concepts have the pure strategy property.
2. All Nash equilibria are rationalizable, all perfect equilibria are perfectly rationalizable and all proper equilibria are properly rationalizable.

Intuitively one would expect perfect and proper rationalizability to be refinements of cautious rationalizability. The following example taken from Pearce (1984) affirms this intuition. Figure 4 gives the payoff matrix of the

	$Y_1$	$Y_2$
$X_1$	10, 10	10, 0
$X_2$	10, 10	0, 0

Fig. 4. A two-player game: G4

normal-form game G4. In G4, action  $X_2$  of player 1 is not perfectly nor properly rationalizable, but it is cautiously rationalizable. However, the next sections make clear that the set of cautiously rationalizable strategy profiles can be either smaller or bigger than the set of strategy profiles obtained by any other refinement of rationalizability.

### 3. New refinements of rationalizability

The basic idea behind refinements of Nash equilibria is that each player makes a mistake with a small probability. The modelling of these mistakes leads to different, but equivalent, definitions of perfect Nash equilibrium. One possibility of modelling these mistakes is to require that each pure strategy is played with at least some minimum probability. Next one considers the Nash equilibria of the resulting perturbed game. A perfect Nash equilibrium is obtained as a limit point of the Nash equilibria of the perturbed games with the probabilities converging to zero. This is exactly the line of reasoning that has been followed to define the concept of perfect rationalizability. An equivalent definition for perfect Nash equilibrium, see van Damme (1991) Theorem 2.2.5, is obtained by considering the notion of an  $\varepsilon$ -perfect equilibrium, i.e. a strategy profile  $c \in \text{int}(M)$  such that for any pure strategy  $s_i \in S_i$  that is not a best response against  $c_{-i}$  it holds that  $c_i(s_i) \leq \varepsilon$ . A perfect Nash equilibrium is a limit point of a sequence of  $\varepsilon$ -perfect Nash equilibria with  $\varepsilon$  converging to zero. Recall that a related modelling of errors, which imposes somewhat more rationality, is taken to define proper Nash equilibria and to define proper rationalizability. It is therefore natural to consider the idea of  $\varepsilon$ -perfection for rationalizability and to ask the question if a refinement of rationalizability is obtained that is equivalent to perfect rationalizability. We call the newly proposed refinement weakly perfect rationalizability.

Given some  $\varepsilon > 0$ , a player  $i$  satisfies the  $\varepsilon$ -perfect trembling condition if, given her conjecture  $c_{-i} \in \text{int}(M_{-i})$ , she plays a completely mixed strategy  $c_i \in \text{int}(M_i)$  such that for any pure strategy  $s_i \in S_i$  that is not a best response against  $c_{-i}$  it holds that  $c_i(s_i) \leq \varepsilon$ . Like Schuhmacher (1995) did for proper rationalizability, one can show that common knowledge among the players of the  $\varepsilon$ -perfect trembling condition implies that every player plays a strategy which survives the following procedure.

**Definition 6.** *Let  $\varepsilon > 0$  be given and let  $D^0(\varepsilon) = \text{int}(M)$ . For  $k \geq 1$ ,  $D^k(\varepsilon) = \prod_{i \in I} D_i^k(\varepsilon)$  is inductively defined as follows:  $c_i$  belongs to  $D_i^k(\varepsilon)$  if  $c_i \in \text{int}(M_i)$  and there is a  $c_{-i} \in \text{ch}(D_{-i}^{k-1}(\varepsilon))$  such that  $c_i(s_i) > \varepsilon$  implies that  $s_i$  is a best response against  $c_{-i}$  within  $S_i$ . The set  $D^\infty(\varepsilon) = \lim_{k \rightarrow \infty} D^k(\varepsilon)$  is the set of  $\varepsilon$ -weakly perfectly rationalizable strategy profiles and  $D^\infty = \lim_{\varepsilon \rightarrow 0^+} D^\infty(\varepsilon)$  the set of weakly perfectly rationalizable strategy profiles.*

Unlike the perfect rationalizability concept, a player is not required to optimize against her conjecture subject to an explicit constraint on minimum probabilities in the weakly perfect rationalizability concept. Instead her conjecture must put weight less than  $\varepsilon$  on strategies that are not best responses.

Reconsider the two-player normal-form game G2 (see Figure 2). Remember that the seemingly unreasonable pure strategy profiles  $(X_1, Y_2)$ ,  $(X_2, Y_1)$  and  $(X_2, Y_2)$  are rationalizable. They are not weakly perfectly rationalizable.

Indeed, it is obvious that  $D^1(\varepsilon) = D^k(\varepsilon)$  for all  $k > 1$  and  $D^1(\varepsilon) = \{c \in \text{int}(M) \mid c_1(X_2) \leq \varepsilon \text{ and } c_2(Y_2) \leq \varepsilon\}$ . So it holds that  $D^\infty = \{(X_1, Y_1)\}$ . Just like the concepts of perfect and proper rationalizability, it is easily verified that in general all weakly dominated strategy profiles are eliminated by weakly perfect rationalizability. Game G4 is therefore an example where cautious rationalizability eliminates less strategies than weakly perfect rationalizability.

A third possible, equivalent, definition for perfect Nash equilibria, see van Damme (1991) Theorem 2.2.5, is the following. A perfect Nash equilibrium is a limit point of a sequence of completely mixed strategy profiles with the property that it is a best reply against every element in the sequence. It follows that a perfect Nash equilibrium is a cautious response. In the same way as rationalizability is related to Nash equilibrium, our newly proposed concept of trembling-hand perfect rationalizability (THR) is related to perfect Nash equilibrium using the third definition. Instead of using best responses, we require players to use cautious responses.

Another motivation which leads to the trembling-hand perfect rationalizability concept is obtained by carrying the logic behind cautious rationalizability one step further. This implicates that one wants to consider a solution concept where players eliminate responses that are not cautious in each round. All pure strategies that haven't been deleted yet are considered as possible by the players, and therefore they do not use conjectures that put probability zero on some of these strategies. THR is defined by the following modification of the iterative procedure of Definition 1.

**Definition 7.** Let  $T^0 = M$ . For  $k \geq 1$ ,  $T^k = \prod_{i \in I} T_i^k$  is inductively defined as follows:  $c_i$  belongs to  $T_i^k$  if  $c_i \in T_i^{k-1}$  and there is  $c_{-i} \in \text{int}(\text{ch}(T_{-i}^{k-1}))$  such that  $c_i$  is a best response against  $c_{-i}$  within  $T_i^{k-1}$ . The set  $T^\infty = \lim_{k \rightarrow \infty} T^k$  is the set of trembling-hand perfect rationalizable strategy profiles.

At each step of the iteration, a strategy  $c_i$  of player  $i$  has to be a best response against some conjecture  $c_{-i} \in \text{int}(\text{ch}(T_{-i}^{k-1}))$ . It follows that at each step of the iteration any weakly dominated strategy is deleted.

It is not too difficult to show the following analogue of Theorem 1.

**Theorem 2.** For every normal-form game  $\Gamma(I, S, U)$ :

1. The sets of weakly perfectly and trembling-hand perfectly rationalizable strategy profiles are non-empty and closed. Both concepts have the pure strategy property.
2. All perfect Nash equilibria are weakly perfectly rationalizable.

One of the motivations for Myerson's (1978) properness notion was that perfectness has the drawback that adding strictly dominated strategies may enlarge the set of perfect equilibria. Nevertheless, van Damme (1991) has shown that, for the equilibrium approach, the properness notion may suffer from the same drawback as well. The game G5 in Figure 5 taken from Pearce (1984) is such an example where both the solution concepts of perfect and proper Nash equilibrium fail to eliminate all intuitively unreasonable outcomes. The game G5 has two pure Nash equilibria:  $\{(X_2, Y_1), (X_1, Y_2)\}$ . In fact, these two Nash equilibria are also perfect and proper Nash equilibria. It

	$Y_1$	$Y_2$
$X_1$	1, 1	1, 1
$X_2$	2, -1	-10, -2
$X_3$	0, -2	0, -1

Fig. 5. A two-player game: G5

follows immediately from Theorems 1 and 2 that these Nash equilibria are perfectly, properly, and weakly perfectly rationalizable. When the strictly dominated strategy  $X_3$  of player 1 is removed, then  $(X_2, Y_1)$  remains as the only strategy profile that is perfectly, properly, and weakly perfectly rationalizable. So these solution concepts are vulnerable to adding a strictly dominated strategy. On the other hand, the mixed strategy combination sets resulting from trembling-hand perfect rationalizability do not change when a strictly or weakly dominated strategy is added to a game. It will be eliminated in the first iteration and does not play a role in subsequent iterations.

In many games, trembling-hand perfect rationalizability can rule out implausible strategies that cannot be excluded by proper rationalizability (although in Section 5 we show that even for two-player normal-form games trembling-hand perfect rationalizability is not a refinement of proper rationalizability). In Game G5, once we apply our concept THR, we obtain the following:  $T^1 = \text{ch}(\{(X_2, Y_1), (X_1, Y_2), (X_2, Y_2), (X_1, Y_1)\})$ ;  $T^2 = \text{ch}(\{(X_2, Y_1), (X_1, Y_1)\})$ ;  $T^3 = \{(X_2, Y_1)\}$ . Once player 1 will never play  $X_3$ , player 2's action  $Y_2$  is never a best response against any trembling conjecture which puts weight on  $X_1$  and  $X_2$ . Therefore,  $Y_1$  is the unique trembling-hand perfect rationalizable strategy of player 2. Knowing that player 2's choice is  $Y_1$ , player 1's best response is to play  $X_2$  which is player 1's unique trembling-hand perfect rationalizable strategy. Game G5 shows that sometimes it is possible to eliminate unreasonable strategies by means of trembling-hand perfect rationalizability which cannot be eliminated by perfect rationalizability, proper rationalizability, weakly perfect rationalizability, or even the proper equilibrium concept since the strategy profile  $(X_1, Y_2)$  constitutes a proper equilibrium in Game G5. Also cautious rationalizability leads to the strategy profile  $(X_2, Y_1)$  in game G5,  $C^\infty = \{(X_2, Y_1)\}$ . However, as already noted before, cautious rationalizability does not even always eliminate weakly dominated strategies, recall for instance game G4.

#### 4. General relationships between refinements

Intuitively one would expect that it is possible to give some generally holding relationships between the refinements. Based on the experience gained from equilibrium concepts one would expect that perfect rationalizability and weakly perfect rationalizability coincide and that both are refinements of proper rationalizability. This intuition is reinforced since Theorems 1 and 2 show that the solution given by these concepts includes the strategy profiles selected by the corresponding equilibrium concepts. The examples given so far show that THR might be a refinement of all other rationalizability concepts,

and that there is no general relationship between cautious rationalizability and any of the other refinements. The first generally holding relationship shows that perfect rationalizability implies weakly perfect rationalizability.

**Theorem 3.** *Every perfectly rationalizable strategy profile is weakly perfectly rationalizable.*

*Proof.* Take any strictly positive vector of probabilities  $\mu$  small enough to ensure that each  $M_i(\mu)$  has full dimension. Let  $\varepsilon = \max_{i \in I} \max_{s_i \in S_i} \mu_i(s_i)$ . It suffices to show that  $B_i^k(\mu) \subseteq D_i^k(\varepsilon)$  for all  $k$ . We prove this by induction on  $k$ . For  $k = 0$ , it is obviously true. Now, let  $B_j^{k-1}(\mu) \subseteq D_j^{k-1}(\varepsilon)$  for all  $j$  and let  $c_i \in B_i^k(\mu)$ . Then, there is  $c_{-i} \in \text{ch}(B_{-i}^{k-1}(\mu))$  such that  $c_i$  is a best response to  $c_{-i}$  within  $M_i(\mu)$ . Since  $\text{ch}(B_{-i}^{k-1}(\mu)) \subseteq \text{ch}(D_{-i}^{k-1}(\varepsilon))$  and  $c_i(s'_i) = \mu_i(s'_i) \leq \varepsilon$  if  $c_i(s'_i, c_{-i}) < c_i(s_i, c_{-i})$  it follows that  $c_i \in D_i^k(\varepsilon)$ . ■

In Section 5 we will give an example showing that the converse of Theorem 3 is not necessarily true. There exist games where the set of perfectly rationalizable strategy profiles is a proper subset of the set of weakly perfectly rationalizable ones.

It is also true that proper rationalizability is a refinement of weakly perfect rationalizability. This is shown in the next result.

**Theorem 4.** *Every properly rationalizable strategy profile is weakly perfectly rationalizable.*

*Proof.* Take any  $\varepsilon \in (0, 1)$  and any player  $i \in I$ . It suffices to show that  $A_i^k(\varepsilon) \subseteq D_i^k(\varepsilon)$  for all  $k$ . We prove this by induction on  $k$ . For  $k = 0$ , this is true since  $A_i^0(\varepsilon) = D_i^0(\varepsilon)$ . Now, let  $A_j^{k-1}(\varepsilon) \subseteq D_j^{k-1}(\varepsilon)$  for all  $j$  and let  $c_i \in A_i^k(\varepsilon)$ . Then it is straightforward that  $c_i \in D_i^k(\varepsilon)$ . ■

The converse of Theorem 4 is not true. In game G3 it holds that  $A^\infty$  is a proper subset of  $B^\infty$ , and in Theorem 3 it is shown that always  $B^\infty \subseteq D^\infty$ .

We have already seen that in game G5 trembling-hand perfect rationalizability is a more powerful refinement than perfect and proper rationalizability. Therefore, we might expect that trembling-hand perfect rationalizable strategy profiles are also weakly perfectly rationalizable, and possibly even that they are perfectly or properly rationalizable. The latter statement is shown to be false in Section 5. Theorem 5 shows that the former statement is true for two-player games. We denote the pure strategies in  $T_i^k$  by  $ST_i^k$ , and the pure strategies that are approximately in  $D_i^k(\varepsilon)$  by  $SD_i^k(\varepsilon)$ . So,  $SD_i^k(\varepsilon) = \{s_i \in S_i \mid \exists c_i \in D_i^k(\varepsilon) \text{ with } c_i(s'_i) = \varepsilon, \forall s'_i \neq s_i\}$ .

**Theorem 5.** *For any two-player game in normal-form, every trembling-hand perfect rationalizable strategy profile is weakly perfectly rationalizable.*

*Proof.* Let  $\bar{\varepsilon} = 1/(\max_{i \in I} \# S_i)$ . First we will show by induction on  $k$  that  $ST_i^k \subseteq SD_i^k(\varepsilon)$ , for all  $\varepsilon \in (0, \bar{\varepsilon})$ . It is easily verified that  $ST_i^0 = SD_i^0(\varepsilon) = S_i$ .

Now, let  $ST_j^{k-1} \subseteq SD_j^{k-1}(\varepsilon)$  for all  $j$ . If  $s_i^1 \in ST_i^k$ , then there is  $c_j^1 \in \text{int}(\text{ch}(T_j^{k-1}))$ ,  $j \neq i$ , and for every  $c_i \in T_i^{k-1}$ ,  $U_i(s_i^1, c_j^1) \geq U_i(c_i, c_j^1)$ . Suppose there is  $s_i^* \in S_i \setminus ST_i^{k-1}$  such that  $U_i(s_i^*, c_j^1) > U_i(s_i^1, c_j^1)$ . Let  $S_i^*$  be the set of all

best responses in  $S_i$  to  $c_j^1$ , so  $S_i^* \subseteq S_i \setminus ST_i^{k-1}$ . Let  $l$  be the maximal integer such that  $S_i^* \cap ST_i^{l-1} \neq \emptyset$ . Since  $c_j^1 \in \text{ch}(T_j^{k-1}) \subseteq \text{ch}(T_j^{l-1})$  and each pure strategy in  $S_i^* \cap ST_i^{l-1}$  is a best response to  $c_j^1$ , at least one pure strategy in  $S_i^* \cap ST_i^{l-1}$  is a best response within  $T_i^{l-1}$  to a sufficiently small perturbation of  $c_j^1$  that gives positive weight to each pure strategy in  $T_j^{l-1}$ . Therefore,  $S_i^* \cap ST_i^l \neq \emptyset$ , a contradiction. Consequently,  $U_i(s_i^1, c_j^1) \geq U_i(s_i, c_j^1)$ , for every  $s_i \in S_i$ .<sup>3</sup> Since  $s_i^1 \in ST_i^1$ , there is  $c_j^2 \in \text{int}(M_j)$  such that  $U_i(s_i^1, c_j^2) \geq U_i(s_i, c_j^2)$ , for every  $s_i \in S_i$ . It follows that  $U_i(s_i^1, (1-\varepsilon)c_j^1 + \varepsilon c_j^2) \geq U_i(s_i, (1-\varepsilon)c_j^1 + \varepsilon c_j^2)$ , for every  $s_i \in S_i$ . Moreover,  $(1-\varepsilon)c_j^1 + \varepsilon c_j^2$  is a completely mixed strategy putting weight less than  $\varepsilon$  on each pure strategy in  $S_j \setminus SD_j^{k-1}(\varepsilon)$ ,  $j \neq i$ , where the induction hypothesis is used for the inclusion. It follows that  $(1-\varepsilon)c_j^1 + \varepsilon c_j^2 \in \text{ch}(D_j^{k-1}(\varepsilon))$ . So,  $c_i^1 \in D_i^k(\varepsilon)$  where  $c_i^1(s_i) = \varepsilon$ ,  $\forall s_i \neq s_i^1$ , and hence  $s_i^1 \in SD_i^k(\varepsilon)$ . We have shown that  $ST_i^k \subseteq SD_i^k(\varepsilon)$ .

Since the sets  $T_i^k$  and  $D_i^k(\varepsilon)$  can only change in the next iteration if the sets  $ST_i^k$  and  $SD_i^k(\varepsilon)$  change, it follows that  $\forall k, l \geq m = \sum_{i \in I} (\# S_i - 1)$ ,  $T_i^k = T_i^l$  and  $D_i^k(\varepsilon) = D_i^l(\varepsilon)$ . If  $c_i' \in T_i^\infty$ , then  $c_i' \in T_i^k$  with  $k \geq m + 1$ , so there is  $c_j^3 \in \text{int}(\text{ch}(T_j^k))$  such that for every  $c_i \in T_i^k$ ,  $U_i(c_i', c_j^3) \geq U_i(c_i, c_j^3)$ . Since  $c_i' \in T_i^1$ , there is  $c_j^4 \in \text{int}(M_j)$  such that  $U_i(c_i', c_j^4) \geq U_i(s_i, c_j^4)$ ,  $\forall s_i \in S_i$ . As in the first part of the proof it follows that  $(1-\varepsilon)c_j^3 + \varepsilon c_j^4 \in \text{ch}(D_j^k(\varepsilon))$  and that  $c_i'$  is a best response against this strategy. So,  $c_i''(\varepsilon) \in D_i^k(\varepsilon) = D_i^\infty(\varepsilon)$  where  $c_i''(\varepsilon)(s_i) = \varepsilon$ , if  $c_i'(s_i) = 0$ , and  $c_i''(\varepsilon)(s_i) = c_i'(s_i)(1 - \#\{s_i' \mid c_i'(s_i') = 0\}\varepsilon)$  if  $c_i'(s_i) \neq 0$ . If  $\varepsilon \rightarrow 0^+$ , then  $c_i''(\varepsilon) \rightarrow c_i'$ , so  $c_i' \in D_i^\infty$ . ■

The proof of Theorem 5 is only valid for the two-player case since it relies on the linearity of  $U_i(s_i, \cdot)$ . Theorem 5 cannot be generalized to three or more player games as is shown by Game G6 (see Figure 6). It is easily seen that  $ST_1^1 = \{X_1, X_2, X_3\}$ ,  $ST_2^1 = \{Y_1, Y_2\}$ , and  $ST_3^1 = \{Z_1, Z_2\}$ . It is not possible in the first iteration to eliminate any pure strategy of player 1, since all strategies of player 1 are equally good against  $(c_2, c_3) = ((1/3, 1/3, 1/3), (1/3, 1/3, 1/3))$ . In the second iteration it is clearly impossible to eliminate any other pure strategy of player 2 or 3. Against  $(c_2, c_3) = ((1/2, 1/2, 0), (1/2, 1/2, 0))$  all pure strategies of player 1 are equally good, so no further

	Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>		Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>		Y <sub>1</sub>	Y <sub>2</sub>	Y <sub>3</sub>
X <sub>1</sub>	2, 1, 1	1, 1, 1	0, 0, 1	X <sub>1</sub>	1, 1, 1	0, 1, 1	0, 0, 1	X <sub>1</sub>	1, 1, 0	1, 1, 0	0, 0, 0
X <sub>2</sub>	0, 1, 1	1, 1, 1	0, 0, 1	X <sub>2</sub>	1, 1, 1	2, 1, 1	0, 0, 1	X <sub>2</sub>	1, 1, 0	1, 1, 0	0, 0, 0
X <sub>3</sub>	2, 1, 1	0, 1, 1	0, 0, 1	X <sub>3</sub>	0, 1, 1	2, 1, 1	0, 0, 1	X <sub>3</sub>	0, 1, 0	0, 1, 0	2, 0, 0
	Z <sub>1</sub>				Z <sub>2</sub>				Z <sub>3</sub>		

Fig. 6. A three-player game: G6

<sup>3</sup> One of the referees had a nice induction argument to show the similar result that if  $c_{-i} \in \text{ch}(T_{-i}^{k-1})$  and  $c_i \in T_i^{k-1}$  is a best response to  $c_{-i}$  in  $T_i^{k-1}$ , then  $c_i$  is also a best response to  $c_{-i}$  in  $M_i$ .

**Table 1.** The payoffs of the pure strategies of player 1

Strategy	Payoff
$X_1$	$2st + (1 - s - \beta)t + s(1 - t - \gamma) + \gamma(1 - \beta)$
$X_2$	$(1 - s - \beta)t + s(1 - t - \gamma) + 2(1 - s - \beta)(1 - t - \gamma) + \gamma(1 - \beta)$
$X_3$	$2st + 2(1 - s - \beta)(1 - t - \gamma) + 2\beta\gamma$

eliminations are possible. Consequently, for every  $k \geq 1$ ,  $ST_1^k = \{X_1, X_2, X_3\}$ ,  $ST_2^k = \{Y_1, Y_2\}$ , and  $ST_3^k = \{Z_1, Z_2\}$ .

Now we consider the weakly perfect rationalizability concept. Let any  $\varepsilon$  smaller than  $1/3$  be given. Obviously, in the first iteration again only the pure strategies  $Y_3$  and  $Z_3$  are eliminated, so  $SD_1^1(\varepsilon) = \{X_1, X_2, X_3\}$ ,  $SD_2^1(\varepsilon) = \{Y_1, Y_2\}$ , and  $SD_3^1(\varepsilon) = \{Z_1, Z_2\}$ . In the second iteration, it is again impossible to eliminate any other pure strategy of player 2 or 3. We show that pure strategy  $X_3$  of player 1 is eliminated in the second iteration, although it is easily seen that  $X_3$  is not weakly dominated by any mixed strategy. Intuitively, compared to strategies  $X_1$  and  $X_2$ , strategy  $X_3$  is good against the conjectures  $(Y_1, Z_1)$ ,  $(Y_2, Z_2)$ , and  $(Y_3, Z_3)$ , but bad against all other pure strategy combinations. If every pure strategy is played with at least a small probability, then the pure strategy combinations against which strategy  $X_3$  is bad will necessarily arise with positive probability. It turns out that against any such conjecture at least one of the pure strategies  $X_1$  and  $X_2$  performs better. Let any  $c_2 \in \text{ch}(D_2^1(\varepsilon))$  and any  $c_3 \in \text{ch}(D_3^1(\varepsilon))$  be given. To simplify notation, let  $s$  and  $t$  denote the probability of the first action of player 2 and player 3, respectively, and  $\beta$  and  $\gamma$  the probability of the third action of player 2 and player 3, respectively, so  $s = c_2(Y_1)$ ,  $t = c_3(Z_1)$ ,  $\beta = c_2(Y_3) \leq \varepsilon$ , and  $\gamma = c_3(Z_3) \leq \varepsilon$ . Let us consider the payoffs of the pure strategies of player 1 (see Table 1).

Pure strategy  $X_3$  is at least as good as pure strategy  $X_1$  if  $t(3 - 3\beta) + s(3 - 3\gamma) + 3\gamma \leq 4st + 5\beta\gamma + 2 - 2\beta$ . So, if,

$$3 - 3\gamma - 4t > 0 \quad \text{and} \quad s \leq \frac{(3\beta - 3)t - 3\gamma - 2\beta + 2 + 5\beta\gamma}{3 - 3\gamma - 4t} \quad (1)$$

or

$$3 - 3\gamma - 4t < 0 \quad \text{and} \quad s \geq \frac{(3\beta - 3)t - 3\gamma - 2\beta + 2 + 5\beta\gamma}{3 - 3\gamma - 4t}.$$

If  $3 - 3\gamma - 4t = 0$ , then  $X_3$  is strictly worse than  $X_1$ . Consider the case  $3 - 3\gamma - 4t < 0$ . It only holds that the right-hand side, i.e. the minimum probability to be put on strategy  $Y_1$ , is less than  $1 - \beta$  if  $t > 1 - 2\beta\gamma/(1 - \beta)$ . But then  $t + \gamma > (1 - \beta + \gamma - 3\beta\gamma)/(1 - \beta) > 1$  since  $\beta < 1/3$ , a contradiction since the sum of  $t$  and  $\gamma$  should be strictly less than 1. So only case (1) remains.

Pure strategy  $X_3$  is at least as good as  $X_2$  if  $t(1 - \beta) + s(1 - \gamma) + \gamma \leq 4st + 3\beta\gamma$ . So, if,

$$1 - \gamma - 4t > 0 \quad \text{and} \quad s \leq \frac{3\beta\gamma - \gamma - (1 - \beta)t}{1 - \gamma - 4t}$$

or

$$1 - \gamma - 4t < 0 \quad \text{and} \quad s \geq \frac{3\beta\gamma - \gamma - (1 - \beta)t}{1 - \gamma - 4t}. \quad (2)$$

If  $1 - \gamma - 4t = 0$ , then  $X_3$  is strictly worse than action  $X_2$ . Consider the case where  $1 - \gamma - 4t > 0$ . It holds that the numerator of the right-hand side is negative (use that  $\beta < 1/3$ ), a contradiction since  $s$  should be positive. So only case (2) remains.

Concluding,  $X_3$  might be a best response of player 1 if

$$1 - \gamma - 4t < 0 < 3 - 3\gamma - 4t$$

and

$$\frac{3\beta\gamma - \gamma - (1 - \beta)t}{1 - \gamma - 4t} \leq s \leq \frac{(3\beta - 3)t - 3\gamma - 2\beta + 2 + 5\beta\gamma}{3 - 3\gamma - 4t}.$$

Next it is shown that the latter inequality can never be satisfied since the first term is always bigger than the third. Now,  $1 - \gamma - 4t < 0 < 3 - 3\gamma - 4t$  and  $(3\beta\gamma - \gamma - (1 - \beta)t)/(1 - \gamma - 4t) \leq ((3\beta - 3)t - 3\gamma - 2\beta + 2 + 5\beta\gamma)/(3 - 3\gamma - 4t)$  implies

$$t^2(4 - 4\beta) + t(4\beta - 4 + 4\gamma - 4\beta\gamma) + 1 - \beta - \gamma - \beta\gamma + 2\beta\gamma^2 \leq 0. \quad (3)$$

The left-hand side of (3) is a quadratic function in  $t$ . Computing “ $b^2 - 4ac$ ” to find the zero points of this function yields  $16\gamma(\gamma - 1)(1 - 4\beta + 3\beta^2)$  which is smaller than 0 (use  $\beta < 1/3$ ). Therefore, the quadratic function in  $t$  has no zero points. By trying any value of the parameters, one sees that the left-hand side of (3) is actually positive everywhere, leading to a contradiction. There are no values of  $s$  and  $t$ , given any  $\beta, \gamma < 1/3$ , for which  $X_3$  is the best response and  $X_3$  can be eliminated. After this no further eliminations are possible. Consequently, for every  $k \geq 2$ ,  $SD_1^k(\varepsilon) = \{X_1, X_2\} \subset ST_1^k = \{X_1, X_2, X_3\}$ ,  $SD_2^k(\varepsilon) = ST_2^k = \{Y_1, Y_2\}$ , and  $SD_3^k(\varepsilon) = ST_3^k = \{Z_1, Z_2\}$ .

## 5. Remaining relationships

### 5.1. Two more examples

The first example, G7, is due to Börgers (1994). Figure 7 gives us the payoff matrix of this two-player normal-form game. In G7, player 1’s pure strategies or actions  $X_1, X_2, X_3$  and player 2’s actions  $Y_1, Y_2, Y_3$  are properly, trembling-hand perfect, and cautiously rationalizable. Meanwhile, only player 1’s actions  $X_1, X_2$  and player 2’s actions  $Y_1, Y_2, Y_3$  are perfectly rationalizable in G7. Perfect rationalizability eliminates pure strategy  $X_3$  in G7. Given both examples G7 and G3, we conclude that there is no relationship between perfect rationalizability and these other refinements (proper, trembling-hand perfect, and cautious rationalizability): perfect rationalizability may be weaker (example G3) or even stronger (example G7).

	$Y_1$	$Y_2$	$Y_3$
$X_1$	3, 0	1, 0	0, 0
$X_2$	0, 0	1, 0	3, 0
$X_3$	2, 0	0, 0	2, 0
$X_4$	0, 3	0, 2	0, 0
$X_5$	0, 0	0, 2	0, 3

Fig. 7. A two-player game: G7

	$Y_1$	$Y_2$
$X_1$	2, 1	1, 1
$X_2$	1, 1	1, 2
$X_3$	0, 1	0, 0

Fig. 8. A two-player game: G8

The second example is the two-player normal-form game G8. Figure 8 gives us the payoff matrix of G8. In G8, proper and cautious rationalizability single out a unique strategy profile:  $(X_1, Y_2)$ . Nevertheless, player 2's action  $Y_1$  is trembling-hand perfect rationalizable:  $T_1^\infty = \{X_1\}$  and  $T_2^\infty = M_2$ . Therefore, there is no relationship between trembling-hand perfect rationalizability and proper or cautious rationalizability: trembling-hand perfect rationalizability may be weaker (Example G8) or stronger (Examples G4 and G5).

### 5.2. The burning money game

Before concluding we briefly consider Ben-Porath and Dekel's (1992) burning money game to get more insight into the consequences of using a particular refinement. This two-stage game is based on an idea of van Damme (1989). In the first stage, player 1 has a choice between burning money (action  $B$ ) and not burning money (action  $N$ ). After this choice is observed, player 1 and 2 play a simultaneous-move game of coordination (actions  $X_1$  or  $X_2$  for player 1 and actions  $Y_1$  or  $Y_2$  for player 2). The corresponding normal-form of this game is given in Figure 9.

For this burning money game, trembling-hand perfect rationalizability singles out a unique strategy profile:  $(NX_1, Y_1 Y_1)$ ; that is, the fact that player 1 could have chosen to burn utility but did not do so ensures that she obtains her most preferred outcome. Indeed, in the game G9, once we apply our concept THR, we obtain the following iterative deletion of pure strategies:  $BX_2 \notin ST_1^1$ ;  $Y_2 Y_1, Y_2 Y_2 \notin ST_2^2$ ;  $BX_2, NX_2 \notin ST_1^3$ ;  $Y_2 Y_1, Y_2 Y_2, Y_1 Y_2 \notin ST_2^4 = \{Y_1 Y_1\}$ ;  $BX_1, BX_2, NX_2 \notin ST_1^5 = \{NX_1\}$ ;  $T^5 = \{(NX_1, Y_1 Y_1)\}$ . Nevertheless,

	$Y_1 Y_1$	$Y_1 Y_2$	$Y_2 Y_1$	$Y_2 Y_2$
$BX_1$	3, 1	3, 1	-2, 0	-2, 0
$BX_2$	-2, 0	-2, 0	-1, 5	-1, 5
$NX_1$	5, 1	0, 0	5, 1	0, 0
$NX_2$	0, 0	1, 5	0, 0	1, 5

Fig. 9. Ben-Porath and Dekel’s burning money game: G9

player 1’s action  $BX_1$  (where player 1 burns money) is properly rationalizable. Indeed,  $A^1(\varepsilon)$  is such that for all  $(c_1, c_2) \in A_1^1(\varepsilon) \times A_2^1(\varepsilon) : c_1(BX_2) \leq \varepsilon c_1(NX_1)$  and  $c_1(BX_2) \leq \varepsilon c_1(NX_2)$ . Given these restrictions, for each pure strategy of player 2 there exists a conjecture  $c_1 \in A_1^1(\varepsilon)$  such that it is a best response against  $c_1$ . Indeed, for all  $c_1 \in A_1^1(\varepsilon)$ , player 2’s expected payoffs are:  $U_2(c_1, Y_1 Y_1) = c_1(BX_1) + c_1(NX_1)$ ;  $U_2(c_1, Y_1 Y_2) = c_1(BX_1) + 5 c_1(NX_2)$ ;  $U_2(c_1, Y_2 Y_1) = 5 c_1(BX_2) + c_1(NX_1) \leq (1 + 5\varepsilon) c_1(NX_1)$ ;  $U_2(c_1, Y_2 Y_2) = 5 c_1(BX_2) + 5 c_1(NX_2) \leq (5 + 5\varepsilon) c_1(NX_2)$ . For example, for all  $\varepsilon \in (0, 1)$ , each pure strategy of player 2 is a best response against the conjecture  $c_1 \in A_1^1(\varepsilon)$  defined by  $c_1(BX_1) = \varepsilon / (6(1 + (1/5)\varepsilon))$ ,  $c_1(BX_2) = ((1/5)\varepsilon) / (6(1 + (1/5)\varepsilon))$ ,  $c_1(NX_1) = 5 / (6(1 + (1/5)\varepsilon))$ ,  $c_1(NX_2) = 1 / (6(1 + (1/5)\varepsilon))$ . For each pure strategy of player 1 belonging to  $\{BX_1, NX_1, NX_2\}$  there exists a conjecture  $c_2 \in A_2^2(\varepsilon)$  such that it is a best response against  $c_2$ . For example, each pure strategy belonging to  $\{BX_1, NX_1, NX_2\}$  is a best response against the conjecture  $c_2 \in A_2^2(\varepsilon)$  defined by  $c_2(Y_1 Y_1) = 1/12$ ,  $c_2(Y_1 Y_2) = 29/60$ ,  $c_2(Y_2 Y_1) = 1/12$ ,  $c_2(Y_2 Y_2) = 7/20$ . Then, the sets of properly rationalizable strategies are the limit sets  $A_1^\infty \supset \{BX_1, NX_1, NX_2\}$  and  $A_2^\infty \supset \{Y_1 Y_1, Y_1 Y_2, Y_2 Y_1, Y_2 Y_2\}$ ; only player 1’s pure strategy  $BX_2$  does not belong to  $A_1^\infty$ . Note that  $(NX_1, Y_1 Y_1)$  is also the unique cautiously rationalizable

Table 2. The (no)-relationships between the refinements

$\subseteq$	Perfect	Proper	Cautious	Weakly Perfect	Trembling-Hand
Perfect	×	Ex. G7 Fig. 7	Ex. G4 Fig. 4	Theorem 3	Ex. G7 Fig. 7
Proper	Ex. G3 Fig. 3	×	Ex. G4 Fig. 4	Theorem 4	Ex. G8 Fig. 8
Cautious	Ex. G5 Fig. 5	Ex. G5 Fig. 5	×	Ex. G5 Fig. 5	Ex. G8 Fig. 8
Weakly Perfect	×	×	Ex. G4 Fig. 4	×	Ex. G6 Fig. 6
Trembling-Hand	Ex. G5 Fig. 5	Ex. G5 Fig. 5	Ex. G4 Fig. 4	Theorem 5 (2-pers.)	×

strategy profile, with  $(5, 1)$  as the resulting payoffs. Therefore, trembling-hand perfect and cautious rationalizability single out the outcome of forward induction [see Ben-Porath and Dekel (1992), Hammond (1993), van Damme (1989)], while proper rationalizability (or weakly perfect rationalizability or perfect rationalizability) does not.

### 5.3. Conclusion

We conclude by summarizing the (no)-relationships between the refinements of rationalizability for normal-form games (see Table 2). The interpretation of an entry in the matrix is that the solution provided by a concept in the row is a subset of the solution provided by a concept in the column.

## References

- [1] Ben-Porath E, Dekel E (1992) Signaling future actions and the potential for sacrifice. *Journal of Economic Theory* 57:36–51
- [2] Bernheim D (1984) Rationalizable strategic behavior. *Econometrica* 52:1007–1028
- [3] Börgers T (1994) Weak dominance and approximate common knowledge. *Journal of Economic Theory* 64:265–276
- [4] Brandenburger A, Dekel E (1987) Rationalizability and correlated equilibria. *Econometrica* 55:1391–1402
- [5] Hammond P (1993) Aspects of rationalizable behavior. In: Binmore K, Kirman A, Tani P (eds) *Frontiers of game theory*, MIT Press, pp 277–305
- [6] Myerson RB (1978) Refinements of the Nash equilibrium concept. *International Journal of Game Theory* 7:73–80
- [7] Pearce DG (1984) Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52:1029–1050
- [8] Schuhmacher F (1995) Proper rationalizability and backward induction. Mimeo, University of Bonn, Bonn
- [9] Selten R (1975) Re-examination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4:25–55
- [10] Van Damme E (1989) Stable equilibria and forward induction. *Journal of Economic Theory* 48:476–496
- [11] Van Damme E (1991) *Stability and perfection of Nash equilibria* Second Edition, Springer-Verlag, Berlin, New York