Games with Imperfectly Observable Commitment*

Eric van Damme[†] Sjaak Hurkens[†]

CentER for Economic Research
Tilburg University
P.O. Box 90153
5000 LE Tilburg, The Netherlands.

July 1994, Revised June 1995

Abstract

In Bagwell (1995) it is claimed that, in models of commitment, "the first-mover advantage is eliminated when there is a *slight* amount of noise associated with the observation of the first-mover's selection." We show that the validity of this claim depends crucially on the restriction to pure strategy equilibria. The game analyzed by Bagwell always has a mixed equilibrium that is close to the Stackelberg equilibrium when the noise is small. Furthermore, an equilibrium selection theory, that combines elements from the theory of Harsanyi and Selten (1988) with elements from the theory of Harsanyi (1995), actually selects this "noisy Stackelberg equilibrium."

Journal of Economic Literature Classification Number: C72.

†The research of Sjaak Hurkens was sponsored by the Foundation for the Promotion of Research in Economic Sciences, which is part of the Netherlands Organization for Scientific Research (NWO). The authors thank an anonymous referee, V. Bhaskar, Werner Güth and seminar audiences at Berlin (Humboldt), Lausanne, Louvain and Tilburg for their constructive criticisms.

1 Introduction

One of the most important insights in game theory is that the power to commit oneself may confer a strategic advantage: it may be beneficial to constrain one's own behavior in order to induce others to behave in a way that is favorable to oneself. One possibility to commit oneself is to move early: to preempt the others by choosing and communicating the (irreversible) action that one takes before the rivals take their actions. This idea dates back at least to Von Stackelberg (1934) who demonstrated the existence of a "first-mover advantage" in a quantity-setting duopoly. Schelling's (1960) classic *The Strategy of Conflict* generalized Von Stackelberg's initial insight in several dimensions by describing richer commitment tactics as well as by illustrating the ubiquity of the phenomenon that in independent decision situations weakness confers strength, that power may result from the power to bind oneself.

Schelling already pointed out that, for a commitment to an action to be credible, the commitment must be irreversible. At least reneging should be sufficiently costly. Schelling also stressed that the efficacy of commitment depends on the communication structure of the game. If the opponent is unavailable for messages, or can destroy all communication channels before any communication takes place, being able to commit oneself is of no value. Hence, commitment can be beneficial only if the communication channel is sufficiently reliable.

Just how important this reliability of the communication channel is has recently been shown in Bagwell (1995). Bagwell shows that it is essential that there are no ambiguities, that there are no misunderstandings about the action to which the player committed himself. In fact, Bagwell claims that "the first-mover advantage is eliminated when there is even a *slight* amount of noise associated with the observation of the first-mover's selection" (Bagwell (1995, Abstract), emphasis in original), and he concludes that "the many predictions derived from models with commitment may require reconsideration" since "the requirement of perfect observability is quite stringent and would seem to be violated in a number of real-world settings to which popular commitment models are thought to apply" (Bagwell (1995, p. 278)).

Bagwell bases the above claim and conclusion on his result that, if there is some noise, a pure strategy Nash equilibrium outcome of the game in which one of the players can commit must be a Nash equilibrium outcome of the game in which this commitment possibility is absent. The intuition for this striking result can be easily conveyed. Let $g = (A_1, A_2, u_1, u_2)$ be a 2-person normal form game and consider the sequential move game with player 1 moving first. In contrast to the usual Stackelberg model, assume, however, that player 2 is only imperfectly informed about player 1's commitment. Specifically, if player 1 commits to $a_1 \in A_1$, player 2 receives the signal $a_1' \in A_1$ with probability $p(a_1' \mid a_1) > 0$. The crucial observation is that this "nonmoving support" assumption implies that player 2's signal is completely uninformative when player 1 commits to a pure action a_1^*: Bayes' rule dictates that 2 believes that 1 played a_1^* no matter what signal he receives. If 2's best response to a_1^* in g is unique (say it is a_2^*), then 2 will respond with a_2^* no matter what message he receives, hence, in order to have an equilibrium in the sequential move game, a_1^* should be a best response against a_2^* as well. Consequently, (a_1^*, a_2^*) must be an equilibrium of g. Note that this argument applies even in the case where player 2 is almost perfectly observed about the commitment, i.e. if $p(a_1 \mid a_1) \approx 1$ for all $a_1 \in A_1$.

Bagwell's result is obviously driven by the specific type of imperfection in the communication technology that he assumes. It is not the case that the commitment sometimes is not communicated, it is rather that the opponent with a small probability receives the wrong message. To put it differently, Bagwell's is a model of errors in perception, rather than of errors in communication. For example, if a seller commits himself to "I do not sell for a price less than \$100", the buyer might interpret this as "I do not sell for less than \$10,000" or as a commitment to "I give the object away for free". Bagwell's "nonmoving support" restriction implies that all such misunderstandings are possible, albeit possibly unlikely. We do not want to enter into the debate about whether, or in which contexts, the assumed communication technology is appropriate. With Bagwell we agree that such discussions are best carried out in the context of specific applications. Nevertheless, we believe that this specific assumption might explain why Bagwell's result appears counterintuitive at first. For example, if communication errors would take the form as suggested in Schelling (1960, p. 149) (i.e. commitments would not necessarily be communicated to the second mover, but if they are communicated, they are communicated without error), then there would not be a lack of robustness of the type that Bagwell notes. The reader can easily verify that in the latter case, as long as the probability that the commitment is received is sufficiently high, a player will commit himself to his Stackelberg strategy. (See Chakravorti and Spiegel (1993) for a formal analysis.)

As Bagwell's claim and conclusion are based on a result concerning pure strategy equilibria, they rely either on the implicit assumption that it is sufficient to look at pure strategy equilibria, or on a belief that a similar result could also be obtained for mixed strategy equilibria. In our opinion, the restriction to pure strategy equilibria is not compelling and the game theory literature has offered no justification for this restriction so far. In addition, as is well known, pure strategy Nash equilibria might fail to exist and existence might be viewed as the most fundamental property that a solution concept should satisfy. Consequently, in this paper we take the position that there is no a priori reason to discriminate against equilibria that are not in pure strategies. We take mixed strategies seriously and investigate what remains of the above claims when these are taken into account.

In the context of a specific 2×2 example, Bagwell already showed that his claim and conclusion depend on the restriction to pure strategy equilibria. Specifically, he showed that "the Stackelberg outcome can be approximately recovered from one mixed-strategy Nash equilibrium for the noisy-leader game when observability is "close to perfect" (Bagwell (1995, p. 277)). We show in this paper that this statement is generally true, i.e. each noisy-leader game has a mixed-strategy equilibrium that generates an outcome that converges to the Stackelberg outcome when the noise vanishes (Proposition 3). We also show that, besides this noisy Stackelberg equilibrium and the pure equilibria uncovered by Bagwell, there may be other equilibria as well.

Bagwell's game, hence, raises the issue of equilibrium selection: If the leader's commitment can only be imperfectly observed, will players coordinate on a pure equilibrium of the simultaneous move game? (and, if they do, on which one?), or will they coordinate on a noisy Stackelberg equilibrium?, or will some completely different equilibrium result? Bagwell does not address this selection issue. He argues that, since the noisy game admits no off-equilibrium path information sets, backward-induction based refinements are not particularly helpful. Consequently, he does not discriminate among Nash equilibria and he concludes that there is certainly no guarantee that a noisy Stackelberg equilibrium will in fact obtain (Bagwell (1995, p. 277)). One aim of this paper is to explicitly address this selection problem and we give arguments for why the noisy Stackelberg equilibrium might be the unique focal one.

A general theory of equilibrium selection has been described in Harsanyi and Selten (1988). An alternative theory has been given in Harsanyi (1995). As we show in this paper, neither theory need select the noisy Stackelberg equilibrium in general. However, as we argue, this is (partly) due to the fact that these theories mix arguments that correspond to the evolutionary (learning) interpretation of the Nash equilibrium concept with arguments that are associated with the rationalistic (one-shot) interpretation of that concept. In our opinion, it is preferable to keep these arguments separate. While discriminating against mixed equilibria may be justified in an evolutionary context, such discrimination most probably is not appropriate in a purely deductive context. We show that in the latter case, the noisy Stackelberg equilibrium may indeed be focal. Specifically, we device a new theory of equilibrium selection, that combines elements from the Harsanyi/Selten theory with elements from the new theory of Harsanyi that does select a noisy Stackelberg equilibrium.

No doubt, some readers may criticize our "mix-and-match"-approach and some of this criticism may be justified. Nevertheless, we wish to argue that our theory has a certain intuitive appeal and that it is no less ad hoc than any of the other theories. Furthermore, at present there is no unanimously accepted theory of equilibrium selection and the various theories that have been proposed are far from perfectly understood. In the absence of convincing axiomatizations — of which we are far removed at the present stage — the only way to increase our understanding seems to be to compute various examples, to see the theories in action. The noisy-leader game provides an interesting test bed to compare these various theories and indeed the comparisons are illuminating. Hence, although we would conclude that the main message of this paper is that there is no immediate need to reconsider the literature that applies the idea of a "first-mover advantage", the paper may also be read as an exercise in equilibrium selection.

The remainder of the paper is organized as follows. Section 2 introduces the noisy commitment game. The Nash equilibria of this game are investigated in Section 3. In particular it is proved that the noisy game admits a mixed equilibrium of which the outcome converges to the Stackelberg outcome of the game with perfect observability when the noise vanishes. Section 4 describes our equilibrium selection theory and proves that it selects a noisy Stackelberg equilibrium. Section 5 is devoted to the distinction between evolutionary and eductive theories and provides a partial justification for Bagwell's claim in an evolutionary context. The Sections 6 and 7 are, respectively, devoted to the theories of Harsanyi (1995) and Harsanyi and Selten (1988). Section 8 concludes.

2 The Noisy Commitment Game

Let g be a (finite) 2-person game in strategic form. Since below we will mainly be interested in what happens when the players move sequentially rather than simultaneously, we label the players as L (for leader) and F (for follower). I (resp. J) denotes the set of pure strategies of player L (resp. player F) in g and u_{ij} (resp. v_{ij}) denotes this player's payoff when the strategy pair (i, j) is played. We identify I (resp. J) with the first $|I|$ (resp. $|J|$) positive integers. Throughout this paper we assume that g satisfies the following regularity condition[1]:

$$\text{if } \ (i,j) \neq (k,l), \text{ then } \ u_{ij} \neq u_{kl} \text{ and } \ v_{ij} \neq v_{kl}. \tag{2.1}$$

This assumption implies that F has a unique best response against each pure strategy i of L. This best response will be denoted by b_i and we write

$$u_i = u_{ib_i}. \tag{2.2}$$

Given (2.1), the sequential move game in which L moves before F and in which F is perfectly informed about the pure action that L has chosen has a unique subgame perfect equilibrium. We use b to denote the SPE strategy of F in this game, hence, b responds to i with b_i ($i \in I$). Without further loss of generality we assume that

$$u_1 > \max_{i \neq 1} u_i \tag{2.3}$$

so that the SPE-outcome is $(1, b_1)$ if L's choice is observed perfectly.

In the remainder we focus our attention on the noisy version of the sequential move game in which F is only imperfectly informed about which action has been chosen by L. To that end, let P be a stochastic matrix defined on the state space I. Hence, $P = (p_{ik})_{i,k \in I}$ with $p_{ik} \geq 0$ and $\sum_k p_{ik} = 1$ for all i. The interpretation is that F receives the signal "L played k" with a probability p_{ik} in case L actually plays i. Emphasis will be on the situation where the noise, i.e. the probability of receiving the "wrong" signal is small but positive. Writing P^0 for the identity matrix on I (i.e. $p^0_{ii} = 1$ for all i) we measure the absolute level of the noise by the distance between P and P^0 and we write

$$|P| = \max\{|p_{ik} - p^0_{ik}| : i, k \in I\}. \tag{2.4}$$

We will restrict ourselves to the case where any signal can result from any action, i.e. just as Bagwell, we impose a "nonmoving support" assumption:

$$p_{ik} > 0 \quad \text{for all } i, k \in I. \tag{2.5}$$

Formally then, we consider the extensive form game g^P given by the following rules:

1. player L chooses an action $i \in I$,
2. chance chooses $k \in I$ with probability p_{ik},
3. player F learns k and chooses $j \in J$,
4. player L receives the payoff u_{ij} and F receives v_{ij}.

The game g^P is referred to as a *noisy commitment game.* Note that the messages (the signals that F receives) are payoff irrelevant. Consequently, g^P differs essentially from a "perturbed game" as used in Selten's (1975) perfectness construction. In a Selten-perturbation of the ordinary sequential move game, if L intends to choose i, he trembles with small probability and, hence, he might actually choose k; F then observes k, he chooses a response j and the players' payoffs are u_{kj} and v_{kj} respectively. In an equilibrium of such a "Selten-perturbation", F chooses b_k in response to any signal k, hence, L chooses "1" when the trembles are sufficiently small. We will see that the noisy commitment game is not that straightforward to analyse.

We denote a (behavioral) strategy of player L (resp. F) in g^P by s (resp. f) and we write $\sigma = (s, f)$ for a strategy combination. Hence, s is a probability distribution on I, $s \in \Delta(I)$, and f is a map that assigns a probability distribution on J to each element of I, i.e. $f \in \Delta(J)^I$. We let s_i denote the probability that L chooses i while f_{kj} is the probability that F chooses j in response to the message k. We write $f_k = j$ if $f_{kj} = 1$ and use similar conventions throughout the text. The *outcome* of the strategy pair $\sigma = (s, f)$ in g^P is the probability distribution $z^P = z^P(\sigma)$ that σ induces on $I \times J$. Hence, we have that

$$z^P(\sigma)_{ij} = s_i \sum_{k \in I} p_{ik} f_{kj}. \tag{2.6}$$

Note that z^P may involve nontrivial correlation of the players' actions. Player L's (expected) payoff in g^P is written as $u^P(\sigma)$ and F's payoff is denoted by $v^P(\sigma)$, hence

$$u^P(\sigma) = \mathcal{E}(u \mid z^P(\sigma)), \;\; v^P(\sigma) = \mathcal{E}(v \mid z^P(\sigma)). \tag{2.7}$$

A pair $\sigma = (s, f)$ is a *Nash equilibrium* of g^P if s is a best reply against f and f is a best reply against s. Since there are no unreached information sets in g^P, any Nash equilibrium is a sequential equilibrium. Furthermore, f is a best response against s if and only if, for every message k, f_k is a best response against the posterior beliefs that s induces at k. By Bayes' rule, this posterior belief that F associates to $i \in I$ after having received the message k is given by

$$\mu_{ik}^{P,s} = p_{ik} s_i / \sum_{\alpha \in I} p_{\alpha k} s_\alpha, \tag{2.8}$$

so that, for all s with $s_k > 0$

$$\lim_{|P| \to 0} \mu_{ik}^{P,s} = \begin{cases} 1 & \text{if } \; i = k \\ 0 & \text{if } \; i \neq k \end{cases} \tag{2.9}$$

Hence, if the noise is small and F expects L to choose k with positive probability, then he will attach high probability to the event that L actually played k when he receives the message "k". Assumption (2.1) thus implies that F will respond to k with b_k in this case. Lemma 1 proves a slightly stronger statement.

Lemma 1 . *There exists $\varepsilon^* > 0$ such that for all P with $0 < |P| < \varepsilon^*$, all strategy combinations $\sigma = (s, f)$ and all $i \in I$: If $s_i > \sqrt{|P|}$ and f is a best reply against s in g^P, then $f_i = b_i$.*

Proof. The regularity assumption (2.1) implies that there exists $\delta < 1$ such that for all $i \in I$: If player F assigns at least probability δ to L playing i in g, then b_i is the unique best response of F in g. Let ε^* be such that $(1 + \sqrt{\varepsilon^*})^{-1} \geq \delta$, choose P such that $0 < |P| = \varepsilon < \varepsilon^*$ and let $s \in \Delta(I)$ and $i \in I$ be such that $s_i > \sqrt{\varepsilon}$. Then we obtain from (2.8)

$$\begin{aligned}
\mu_{ii}^{P,s} &\geq \frac{p_{ii}s_i}{\varepsilon(1 - s_i) + p_{ii}s_i} \\
&= [1 + \varepsilon(1 - s_i)/p_{ii}s_i]^{-1} \\
&\geq (1 + \sqrt{\varepsilon})^{-1} \\
&\geq (1 + \sqrt{\varepsilon^*})^{-1} \\
&\geq \delta
\end{aligned}$$

If f is a best reply against s, then f_i is a best reply against the posterior beliefs $\mu_{\cdot i}^{P,s}$ for all i, hence, $f_i = b_i$. □

3 Equilibria in the Noisy Commitment Game

For the sake of completeness we start by stating (and proving) Bagwell's main result.

Proposition 1 *(Bagwell (1995)). The set of pure strategy equilibrium outcomes of g and g^P coincide.*

Proof. Assume (i, j) is a pure strategy Nash equilibrium in g. Then $j = b_i$ and if f is the strategy of F in g^P defined by $f_k = b_i$ $(k \in I)$, then (i, f) is an equilibrium of g^P. It obviously produces the same outcome as (i, j) does. Assume (i, f) is a pure strategy

Nash equilibrium in g^P. Since $\mu_{ik}^{P,i} = 1$ for all k, we must have $f_k = b_i$ for all k. Hence, i is a best reply against b_i in g and (i, b_i) is an equilibrium of g with the same outcome as (i, f). □

Proposition 1 gives a sufficient condition for an outcome to be an equilibrium outcome of the game g^P. We now give a necessary condition for the case where the noise is small. Write

$$\mathcal{N} = \{(i, b_i) : \quad u_i \geq \max_k \min_j u_{kj}\} \tag{3.1}$$

for the set of Nash equilibrium outcomes of the game in which player L's commitment is perfectly observed by F. (Note that because of (2.1) any such Nash equilibrium outcome has to be pure.) The next proposition shows that the Nash equilibrium outcome correspondence of g^P is upper hemi continuous at $P = P^0$.

Proposition 2 *Let z^P be an equilibrium outcome of g^P. If $z = \lim_{|P| \to 0} z^P$, then $z \in \mathcal{N}$.*

Proof. The proof follows from regularity assumption (2.1) and Lemma 1. Let ε^* be as in Lemma 1 and for P with $0 < |P| < \varepsilon^*$, let (s^P, f^P) be an equilibrium of g^P with outcome z^P. Assume the limit outcome z to exist. If $i \neq k, s_i^P > \sqrt{|P|}$ and $s_k^P > \sqrt{|P|}$, then $f_i^P = b_i$ and $f_k^P = b_k$, hence

$$\lim_{|P| \to 0} u^P(i, f^P) = u_i, \quad \lim_{|P| \to 0} u^P(k, f^P) = u_k.$$

But (2.1) implies that $u_i \neq u_k$, hence, either i or k is not a best response when $|P|$ is small. Consequently, for $|P|$ sufficiently small, $s_i^P s_k^P = 0$ and the contradiction shows that there is at most one $i \in I$ with $s_i^P > \sqrt{|P|}$. It follows that $z = (i, b_i)$ for this particular value of i. It is obvious that the inequalities in (3.1) must be satisfied. If there would exist $k \neq i$ with $u_i < \min_j u_{kj}$, then L would strictly prefer choosing k above choosing i in g^P for sufficiently small $|P|$. □

Proposition 2 implies that, if g has only mixed equilibria, the equilibrium outcomes of g are disjoint from the limit equilibrium outcomes of the noisy commitment game when the noise vanishes. This shows that a result similar to Proposition 1 cannot be proved for a "satisfactory" solution concept, i.e. for a refinement of the Nash equilibrium concept that generates a nonempty set of solutions for every game.

The game of Figure 1 may show that not any Nash equilibrium outcome of the game with perfect observability can be approximated by Nash equilibrium outcomes of games with slight noise: the equilibrium correspondence is not lower hemi continuous. It is easily checked that (B, W) is a Nash outcome of the non-noisy game: It is optimal for L to commit to B if F responds to T with E. However, since W is a dominant strategy for F in g, noise forces F to choose W in response to any signal. Consequently, only (T, W) can be approximated by equilibrium outcomes of noisy games. (More generally, g^P has a unique equilibrium whenever F has a dominant strategy in g.)

	W	E
T	3,3	0,0
B	2,2	1,1

Figure 1.

Proposition 3 is the main result of this section. It shows that at least one equilibrium outcome of the game with perfect observability, viz. the Stackelberg outcome, is approximated by equilibrium outcomes of the noisy commitment game when the noise vanishes. We will refer to such an approximating equilibrium as a noisy Stackelberg equilibrium.

Proposition 3 . *The game g^P has an equilibrium $\sigma^P = (s^P, f^P)$ with an outcome z^P that converges to $(1, b_1)$ as $|P| \to 0$.*

Proof. Let ε^* be as in Lemma 1 and let $|P| < \varepsilon^*$. Consider the reduced strategic form $\bar{g}^P$ that results from the strategic form of g^P by eliminating all pure strategies of F that do not prescribe to play b_1 after the signal "1". In this reduced game, player L's expected payoff resulting from playing "1" is approximately u_1 no matter what F plays. Let $\sigma^P = (s^P, f^P)$ be an equilibrium of $\bar{g}^P$. If $s_i^P > \sqrt{|P|}$ for some $i \neq 1$, then $f_i^P = b_i$ (Lemma 1) and L's payoff resulting from "i" is approximately u_i. In this case, $u^P(i, f^P) < u^P(1, f^P)$, so that player L wants to choose i with probability zero. The contradiction shows that

$$s_i^P \leq \sqrt{|P|} \text{ for all } i \neq 1, \tag{3.2}$$

hence, $s_1^P \to 1$ as $|P| \to 0$. Lemma 1 thus implies that at the signal "1" only b_1 is a best response of player F, hence, that σ^P is an equilibrium of g^P. Obviously, the outcome z^P of σ^P converges to $(1, b_1)$ as $|P| \to 0$. □

Just as the necessary condition for limit equilibrium outcomes (Proposition 2) is not sufficient (Figure 1), the sufficient conditions (Propositions 1 and 3) are not necessary: Also outcomes that are not pure Nash equilibria, nor Stackelberg equilibria of g may be approximated. Consider the game of Figure 2 in which L has M as a dominant strategy, so that (M, C) is the unique Nash equilibrium. The Stackelberg equilibrium is (T, W). Consider the noisy commitment game with uniform noise, i.e. $p_{ij} = \varepsilon$ if $i \neq j$ and $p_{ii} = 1 - 2\varepsilon$. It is easily seen that the following strategy combination is an equilibrium of this game: Player L commits to M with probability $\frac{3\varepsilon}{1+\varepsilon}$ and to B with the remaining probability $\frac{1-2\varepsilon}{1+\varepsilon}$; player F responds to signals T and B with E, after signal M he plays C with probability $\frac{2}{4-11\varepsilon}$ and E with probability $\frac{2-11\varepsilon}{4-11\varepsilon}$. The corresponding limit outcome is (B, E).

	W	C	E
T	4,4	0,0	0,0
M	5,0	1,1	5,0
B	0,0	0,0	3,3

Figure 2.

We will not attempt to describe exactly which outcomes can be obtained as limits of equilibrium outcomes of the noisy game as the noise tends to zero. Rather we conclude from the Propositions 1 and 3 that typically there exist multiple limits and, hence, that there exists an equilibrium selection problem. We will address this selection problem directly and we will propose a theory according to which a noisy Stackelberg equilibrium is uniquely focal. Our theory incorporates elements from the theory proposed by Harsanyi and Selten (1988) as well as elements from the theory proposed in Harsanyi (1995). However, our theory differs from both its ancestors and it may select different outcomes. In particular, neither the Harsanyi/Selten theory nor the new theory of Harsanyi need to select a noisy Stackelberg equilibrium.

4 Selection of the Stackelberg Equilibrium

In this section we describe a general theory of equilibrium selection and show that this theory selects a noisy Stackelberg equilibrium in g^P when the noise P is small. Hence, when players behave in accordance to this theory, they will consider the Stackelberg equilibrium to be uniquely focal. The theory is simple and is based on the seminal equilibrium selection theories from Harsanyi and Selten (1988) and Harsanyi (1995). In particular, our theory assumes that the players use the tracing procedure from Harsanyi and Selten (1988) to convert their initial (disequilibrium) beliefs into an equilibrium of the game and that players construct their prior beliefs by incorporating the structural incentives to use the various pure strategies, as suggested in Harsanyi (1995).

We do not want to argue that this hybrid theory is necessarily superior to any of its ancestors. It certainly is simpler than the original Harsanyi/Selten theory: The (somewhat cumbersome) tracing procedure has to be applied only once and intransitivities are avoided. In other words, the new theory makes a multilateral risk comparison, while the Harsanyi/Selten theory relies on pairwise comparisons. Furthermore, the hybrid theory is not completely ad hoc: Many of the arguments that Harsanyi/Selten use to defend their theory can be used to defend it. In the absence of axiomatic underpinnings of any of these theories, discussions about their reasonableness should not be solely conducted in the abstract: One has to see the theories in action in order to make informed judgements about their viability. Consequently, we will also attempt to apply the original theories of Harsanyi/Selten and of Harsanyi to the noisy commitment game. We will see that the first theory is difficult to work with and that the second theory yields ambiguous results.

4.1 A Theory of Equilibrium Selection

To describe the general theory it is convenient to simplify notation somewhat. Throughout this subsection we assume an arbitrary 2-person game $\gamma = < A_1, A_2, u_1, u_2 >$ as given. We use subscripts m and n to denote players. Whenever m and n appear in the same expression it is understood that $n \neq m$. We denote a generic pure strategy of player n by a_n and a mixed strategy by s_n. $B_n(s_m)$ denotes the set of all pure best replies of player n against the mixed strategy s_m of player m. The first main building block in our theory is the tracing procedure, which is a model of the players' reasoning process. Starting from given prior expectations (i.e. a mixed strategy pair), it gradually adjusts players' plans and expectations until they are in equilibrium. We only describe the mechanics of this procedure. For the motivation and heuristic description of the process we refer to Harsanyi (1975) and Harsanyi and Selten (1988).

Let p be a mixed strategy combination in γ, which is interpreted as the players' prior expectations. Hence, a priori player n believes that m will play a_m with probability $p_m(a_m)$. For $t \in [0, 1]$ consider the strategic form game $\gamma^{t,p}$ defined by

$$u_n^{t,p}(a) = tu_n(a) + (1-t)u_n(a_n, p_m). \tag{4.1}$$

Hence, for $t = 1$ this game coincides with γ, while for $t = 0$ we have a trivial game in which each player's payoff depends only on his prior expectations. Write $\Gamma(p)$ for the graph of the equilibrium correspondence, i.e.

$$\Gamma(p) = \{(t; s) : t \in [0,1], s \text{ is an equilibrium of } \gamma^{t,p}\}. \tag{4.2}$$

It can be shown that in nondegenerate cases this graph $\Gamma(p)$ contains a unique distinguished curve that connects the unique equilibrium of $\gamma^{0,p}$ to an equilibrium $s(p)$ of γ. (See Schanuel et al. (1991) for details. The interested reader may gain intuition by doing some computations for a 2×2 game. In this case, there generically is a unique curve connecting the equilibrium at $t = 0$ with an equilibrium at $t = 1$.) The (linear) *tracing procedure* consists of following this curve until its endpoint, and the endpoint $T(p) = s(p)$ is called the linear trace of p in γ. The interpretation is that players eventually reason themselves to the equilibrium $T(p)$ if they start from the prior p. Formally, the tracing procedure thus is a map T that maps each mixed strategy vector p of γ into an equilibrium $T(p)$ of γ.

It remains to specify how the prior should be constructed. We propose to follow Harsanyi (1995) and to construct the prior on the basis of the players' structural incentives to use their various pure strategies. The latter are related to the sizes of the stability regions of these strategies. Specifically, for a pure strategy a_n of player n define the *stability set* of a_n as the set $S_m(a_n)$ of all mixed strategies of player m in γ against which a_n is a best reply, i.e.

$$S_m(a_n) = \{s_m \in S_m : \ a_n \in B_n(s_m)\} \tag{4.3}$$

At first it seems natural to measure the structural incentives of a pure strategy a_n by the

Lebesgue measure of $S_m(a_n)$, but Harsanyi (1995) shows that this definition would violate certain desirable properties. To circumvent these, Harsanyi first transforms the strategy simplex S_m by the so-called inversion mapping ω_m and he then takes the Lebesgue measure of the transformed set. Formally, ω_m is the mapping from the interior of S_m to the interior of S_m that maps s_m into $\bar{s}_m$ defined by

$$\bar{s}_m(a_m) = s_m^{-1}(a_m) / \sum_{a \in A_m} s_m^{-1}(a). \tag{4.4}$$

Hence, Harsanyi measures the structural incentive of player n to use the pure strategy $a_n \in A_n$ by the number[2]

$$\psi_n(a_n) = \lambda(\omega_m(S_m(a_n))) \tag{4.5}$$

where λ denotes Lebesgue measure. We propose to use the mixed strategy vector determined by (4.5) as the players' prior assessment about how the game will be played and to use the tracing procedure to convert this prior into an equilibrium of the game. Hence, formally we suggest to take as the solution of the game γ the equilibrium $T(\psi)$ where $\psi = (\psi_1, \psi_2)$ is as in (4.5).

4.2 The Main Result

In this subsection we show that the above theory selects an equilibrium outcome that is close to the Stackelberg outcome of g when the noise in g^P is small. The intuition is simple. If L's commitment is observed perfectly, then the strategy b of player F is a best response against all possible actions of L. While this is no longer true if there is a slight amount of noise (i.e. $P \neq P^0$), it is still quite likely that b is a best response. Specifically, as Lemma 1 has shown, if $|P| < \varepsilon^*$, then b is the unique best response against all s with $s_k > \sqrt{|P|}$ for all k. Hence, the stability set of the strategy b of player F converges to the entire strategy simplex of player L as $|P| \to 0$ and the stability set of any other pure strategy converges to the empty set. It thus follows that the prior of player L, as constructed by (4.5) puts almost all weight on the strategy b of player F when $|P|$ is small. Consequently, L will be tempted to play his Stackelberg leader strategy "1". On the other hand, Lemma 1 also implies that, with player F's prior constructed in a similar way, F is tempted to respond with b_i to any pure strategy i for which L's structural incentive is strictly positive. In particular, (2.1) implies that F initially plays b_1 in response to L's Stackelberg action "1". It is not too surprising that, if players' initial actions are the Stackelberg equilibrium actions, the tracing procedure indeed forces the players to coordinate on a noisy Stackelberg equilibrium. Formally, we have

Proposition 4 . *Let σ^P be the solution of the game g^P according to the theory from Section 4.1 and let z^P be the outcome of this solution. Then $z^P \to (1, b_1)$ as $|P| \to 0$.*

Proof. We have $\sigma^P = T^P(\psi^P)$ where ψ^P describes the players structural incentives in the game g^P, as defined via (4.5) and where T^P denotes the tracing map associated with g^P. It follows directly from the definitions that $\psi_F^P(b) \to 1$ as $|P| \to 0$. Hence, the leader's prior attaches high probability to the follower always playing a "naive" best response to each possible message. Consequently, (2.1) implies that when $|P|$ is small enough

$$u^{P^0}(i, \psi_F^P) < u^{P^0}(1, \psi_F^P) \text{ for all } i \neq 1. \tag{4.6}$$

This condition in turn implies that there exist $\varepsilon > 0$ and $t^* > 0$ such that

$$\text{"1" is strictly dominant for } L \text{ in } g^{P,t,\psi^P} \text{ if } t < t^* \text{ and } |P| < \varepsilon. \tag{4.7}$$

Furthermore, by choosing ε sufficiently small we can guarantee that for all $i \neq 1$:

$$\text{if } |P| < \varepsilon, \text{ then } u^P(i, f) < u^P(1, f) \text{ for all } f \text{ with } f_i = b_i \text{ and } f_1 = b_1. \tag{4.8}$$

We will restrict ourselves to stochastic matrices P with

$$\sqrt{|P|} \leq t^*/2|I|. \tag{4.9}$$

Finally, with ε^* as in Lemma 1, we assume that

$$|P| < \varepsilon^*. \tag{4.10}$$

Let P be such that (4.7) - (4.10) hold and denote by $\sigma^{P,t} = (s^{P,t}, f^{P,t})$ an equilibrium on the distinguished curve in $g^P(\psi^P)$ that connects the unique equilibrium of $g^{P,0,\psi^P}$ with $T^P(\psi^P)$. We claim that

$$s_i^{P,t} < 1/2|I| \text{ for all } i \neq 1 \text{ and all } t. \tag{4.11}$$

Assume, to the contrary, that there exist some $i \neq 1$ and t such that $s_i^{P,t} \geq 1/2|I|$ and let τ be the smallest t for which an equilibrium of this type can be found. Then $\tau \geq t^*$ in view of (4.7). Hence, at $t = \tau$, the total probability that F assigns to L playing i in g^{P,t,ψ^P} is at least $t^*/2|I|$, so that (4.9) and Lemma 1 guarantee that $f_i^{P,\tau} = b_i$. At the same time we have that

$$s_1^{P,\tau} = 1 - \sum_{i \neq 1} s_i^{P,\tau} > 1 - |I|/2|I| = 1/2 \geq \sqrt{|P|}$$

so that $f_1^{P,\tau} = b_1$ by the same argument. But now (4.7) and (4.8) imply that

$$u^{P,\tau,\psi^P}(i, f^{P,\tau}) < u^{P,\tau,\psi^P}(1, f^{P,\tau}),$$

hence, $s_i^{P,\tau} = 0$. The contradiction shows that (4.11) holds. In particular, we have that

$s_1^{P,1} > 1/2$, hence $f_1^{P,1} = b_1$ in view of Lemma 1. Applying Lemma 1 and (4.8) once more we see that, therefore, $s_i^{P,1} \leq \sqrt{|P|}$ for all $i \neq 1$, hence, that

$$\lim_{|P| \to 0} s_1^{P,1} = 1.$$

This completes the proof. □

5 Evolutionary and Eductive Theories

In his unpublished dissertation (Nash (1950)), Nash already discussed two interpretations of the equilibrium concept that he proposed. In the first "mass-action" interpretation, one imagines that the game is played over and over again with players constantly re-drawn from large populations and them accumulating empirical information about the relative attractiveness of their various strategies. A stable rest point of this dynamic process has to be a Nash equilibrium. In the second "rationalistic" interpretation the game is assumed to be played just once and the players are perfectly rational so that they can figure out what each one will do. A rational prediction about how the game will be played has to be a Nash equilibrium since otherwise it is self-defeating. This second interpretation relies essentially on the assumption that each game admits a unique rational solution. The theories of Harsanyi and Selten (1988) and Harsanyi (1995), as well as the theory from Section 4.1 find their origin in the fact that many games admit multiple equilibria, hence, that Nash's second justification is incomplete. These theories attempt to obtain unique solutions by imposing stronger rationality requirements. In the terminology of Binmore (1987), they are eductive theories, and not evolutionary theories.

At the same time the theories of Harsanyi and Selten (1988) and of Harsanyi (1995) incorporate elements that are most (or, perhaps, only) relevant in an evolutionary context. Specifically, these theories are strongly biased in favor of selecting pure (strict) equilibria and they start by eliminating many mixed equilibria. The basic reason for this bias is that mixed equilibria typically have very poor stability properties (cf. Harsanyi and Selten (1988, p. 201) and Harsanyi (1995, footnote 12)). However, one may very well wonder whether such a bias is justified: The stability property may be relevant in an evolutionary context, but it is not clear that it has any relevance in the case where the game is played only once and players rely exclusively on deductive personal reflection in order to figure out what to play. For example, Kohlberg and Mertens' (1986) concept of stable equilibrium attempts to capture strong rationality requirements and each completely mixed equilibrium is "strategically stable" as a singleton. Hence, these theories may be criticized for the fact that they mix arguments that are relevant in an evolutionary context with arguments that are relevant in an eductive context.

Of course, not all games admit pure equilibria, hence, the theories of Harsanyi/Selten and of Harsanyi do not always select strict equilibria. To avoid the non-existence problem, these theories rely on set-valued solution concepts that can be viewed as generalizations of the concept of strict equilibrium. Harsanyi and Selten (1988) start by eliminating equilibria that do not belong to primitive formations and Harsanyi (1995) eliminates all equilibria that are not persistent (Kalai and Samet (1984)). A primitive formation is a minimal set of strategy pairs that is closed with respect to taking best responses[3] and it can be shown that, for generic 2-person games, an equilibrium is persistent if and only if it belongs to a primitive formation. In fact, a primitive formation is nothing else but a persistent retract in that case (Van Damme 1995)). As one of us has shown elsewhere, the stability property captured by persistency may indeed be relevant in an evolutionary context. (Hurkens (1994).)

The game displayed in Figure 3 may show that the restriction to persistent equilibria may eliminate any noisy Stackelberg equilibrium. The game g^P has three equilibria: one corresponds to Proposition 1 (with outcome (B,E)), another corresponds to Proposition 3 (with outcome close to (T,W)), and there is a third mixed strategy equilibrium. Action B (i.e. the dominant strategy of L in g) is used with positive probability in all three equilibria and the unique best response of player F against B in g^P is to always respond with E. Consequently, $\{(B, EE)\}$ is the unique primitive formation in g^P, hence (B, EE) is the unique persistent equilibrium of this game. Therefore, the theories of Harsanyi/Selten and Harsanyi select the pure equilibrium of g as the solution of g^P. These theories confirm Bagwell's claim that slight noise eliminates the commitment power.

	W	E
T	3,3	0,2
B	4,0	1,1

Figure 3.

The argument used in the above example can be generalized. If one accepts persistency as a selection criterion, one is led to the conclusion that in any game that is (strictly) dominance solvable (i.e. the game can be reduced to a single (strict) equilibrium by repeated elimination of strictly dominated strategies) slight noise eliminates the benefits of the leader being able to commit himself. Hence, the following Proposition implies that, for a certain class of games, Bagwell's claim is justified in an evolutionary setting of the game.

Proposition 5 . *If g is (strictly) dominance solvable with outcome (i^*, j^*), then $\{(i^*, \underline{j}^*)\}$ (i.e. L plays i^* and F responds to each signal with j^*) is the unique primitive formation (resp. persistent retract) of g^P, hence, $(i^*, \underline{j}^*)$ is the unique persistent equilibrium of g^P.*

Proof. Since (i^*, j^*) is a strict equilibrium of g, $(i^*, \underline{j}^*)$ is a strict equilibrium of g^P, hence $\{(i^*, \underline{j}^*)\}$ is a primitive formation of g^P. Let $R = R_L \times R_F$ be a formation of g^P. We will show that $(i^*, \underline{j}^*) \in R$.

Let i_1 be an arbitrary pure strategy of L in R_L. If $i_1 = i^*$ we are done, so assume $i_1 \neq i^*$. The unique best response of F against i_1 in g^P is $\underline{b}_{i_1}$ (respond to each message by playing b_{i_1}), hence, $\underline{b}_{i_1} \in R_F$. Let $g^1, g^2, \ldots, g^T$ be the sequence of games that results from g be repeated elimination of all pure strategies that are strictly dominated. Hence, $g^1 = g$, g^T

only contains (i^*, j^*) and, for each player, g^{t+1} only contains those pure strategies from g^t that are not strictly dominated in g^t. Since $i_1 \neq i^*$, there exists $t < T$ such that i_1 is available in g^t but not in g^{t+1}. Then b_{i_1} is available for F in g^τ for $\tau \leq t+1$. Let i_2 be the unique best response of L in g against b_{i_1}. Then i_2 is available in g^τ for $\tau \leq t+1$. Since i_2 is the unique best response against $\underline{b}_{i_1}$ in g^P, we have $i_2 \in R_L$. Furthermore, b_{i_2} is an available strategy for F in g^τ for $\tau \leq t+1$ and $\underline{b}_{i_2} \in R_F$. If $i_2 \neq i^*$, we can repeat the above argument. In this way we generate a sequence $i_1, i_2, \ldots, i_k$ of elements from R_L with $i_k = i^*$. □

In the following sections we will again take a purely deductive perspective, hence, we will not a priori discriminate against equilibria in mixed strategies. As a consequence, we will not restrict attention to persistent equilibria and we will consider appropriate modifications of the theories of Harsanyi and Harsanyi/Selten. We will see that these modifications might, but need not, select a noisy Stackelberg equilibrium.

6 Harsanyi's (1995) Theory

The basic assumption underlying Harsanyi's (1995) theory is that the probability that a player n will use one of his eligible strategies will be proportional to the structural incentive to use this strategy. The structural incentive of a pure strategy is defined as in (4.5) and the structural incentive of a mixed strategy s_n of player n is given by

$$\psi(s_n) = \sum_{a_n \in A_n} s_n(a_n)\psi(a_n). \tag{6.1}$$

(Note that for a mixed strategy it is not meaningful to define the structural incentive as in (4.5): the number is generically zero.) The prior probability of a mixed strategy pair s is just the product of the individual probabilities

$$\psi(s) = \psi_1(s_1)\psi_2(s_2) \tag{6.2}$$

and, whenever there is one eligible equilibrium that has the highest probability, Harsanyi selects this equilibrium as the solution. Formally, Harsanyi defines an equilibrium to be *eligible* if it is both proper (Myerson (1978)) and persistent (Kalai and Samet (1984)). Writing E for the set of eligible equilibria, Harsanyi defines $s^* \in E$ to be the solution of the game if $\psi(s^*) > \psi(s)$ for all $s \in E \backslash \{s^*\}$.

In the remainder of this section we show that, even in the case where the Stackelberg equilibrium is a strict equilibrium of g^P and, hence, satisfies Harsanyi's eligibility criteria, Harsanyi's theory need not select it. Secondly, even the modified Harsanyi theory — that does not impose the eligibility criterion — need not select the Stackelberg equilibrium. To demonstrate these claims, consider the game from Figure 4, in which K is some real positive number.

	W	E
T	2,1	0,0
B	0,0	1,K

	WW	WE	EE
T	2,1	$2(1-\varepsilon)$,$1-\varepsilon$	0,0
B	0,0	$1-\varepsilon$,$K(1-\varepsilon)$	1,K

Figure 4.

The game g from the left panel of Figure 4 is a unanimity game with Stackelberg outcome (T, W). The panel on the right displays (a reduced form of) the game g^P where P involves uniform noise ($p_{ij} = \varepsilon$ if $i \neq j$). We have eliminated the strategy EW for player F in g^P (i.e. the strategy in which F responds to T by E and to B by W) since this is a strictly dominated strategy. Harsanyi indeed suggests to eliminate such strategies before computing the players' structural incentives. The game g^P has three equilibria (T, WW), (B, EE) and a mixed equilibrium. In the latter, player L plays $(K\varepsilon, 1-\varepsilon)/(K\varepsilon + 1 - \varepsilon)$, while F plays $(0, 1, 1-\varepsilon)/(2-\varepsilon)$. The outcome of this mixed equilibrium converges to the pure equilibrium (B, E) as ε tends to zero. Only the two pure equilibria satisfy Harsanyi's eligibility criteria, hence, to compute the Harsanyi solution of the game, we have to compare the prior probabilities of these equilibria.

Note that the structural incentives for player L to use any of his pure strategies are independent of K since they depend only on player L's own payoff matrix. Furthermore, since in the limit (as ε tends to zero) both stability regions have a nonempty interior, the structural incentive of T as well as of B remains bounded away from zero. The structural incentives of player F are simple to calculate since, in the 1-dimensional case, the inversion mapping is measure preserving. Hence, the prior probability of a strategy is just the Lebesgue measure of the stability set of that strategy. Straightforward computations show that

$$\psi_F^\varepsilon(WW) = \varepsilon/(K - K\varepsilon + \varepsilon) \tag{6.3}$$

and

$$\psi_F^\varepsilon(EE) = K\varepsilon/(K\varepsilon + 1 - \varepsilon). \tag{6.4}$$

Hence

$$\lim_{\varepsilon \downarrow 0} \psi_F^\varepsilon(EE)/\psi_F^\varepsilon(WW) = K^2. \tag{6.5}$$

It follows that, if K is sufficiently large

$$\lim_{\varepsilon \downarrow 0} \psi^\varepsilon(T, WW) < \lim_{\varepsilon \downarrow 0} \psi^\varepsilon(B, EE) \tag{6.6}$$

so that Harsanyi's theory selects the equilibrium (B, EE) in that case. For large values of K, Harsanyi's theory does not select the Stackelberg equilibrium.

Note that the quantities in (6.3) and (6.4) converge to 0 as ε tends to 0 since player F's structural incentive for WE converges to 1: In the limit, WE is weakly dominant. The structural incentive of F to use WE plays no role in the comparison between (T, WW) and (B, EE) since none of these equilibria uses the strategy WE. However, this strategy is used with positive probability in the unique mixed equilibrium of g^P. This mixed equilibrium, hence, has a strictly positive structural incentive and Harsanyi's theory would have selected it had it not been eliminated by the persistency requirement. As argued above, the persistency requirement might be out of place in a deductive theory. However, the modified theory that drops this requirement also does not select the Stackelberg equilibrium: the mixed equilibrium outcome converges to (B, E) as the noise vanishes.

7 The Harsanyi/Selten (1988) Theory

Harsanyi and Selten determine the solution of a game γ by applying a recursive procedure, see the flowchart on p. 222 of their book. Starting from a set of initial candidates, they successively eliminate candidates that are either payoff dominated or risk dominated until, eventually, only one equilibrium is left. The first candidate set contains exactly one equilibrium from each primitive formation. Specifically, for a primitive formation A', define $p(A')$, the *centroid* of A', as that mixed strategy pair in which each player i chooses each element from A'_i with equal probability. The first candidate set is the set of outcomes that the tracing procedure yields when applied to these centroids, hence, it is the finite set $\{T((A')) : A' \text{ is a primitive formation}\}$.

The first candidate set is refined by using dominance criteria. The theory incorporates both payoff dominance and risk dominance and the former is considered to be more important. Formally, equilibrium s is said to *dominate* equilibrium s' if either s payoff dominates s' ($u_n(s) > u_n(s')$ for $n = 1, 2$), or s' does not payoff dominate s and s risk dominates s'. Risk dominance is defined by means of the tracing procedure. An equilibrium s is said to risk dominate an equilibrium s' if the tracing procedure, when started at a certain (bicentric) prior $p(s, s')$ ends up at the equilibrium s. Harsanyi/Selten construct the bicentric prior in the following way. Each player n will initially assume that his opponent m already knows which of the two, s or s', is the solution. Player n will assign a subjective probability z_n to the solution being s (and, hence, to m playing s_m) and he will assign the complementary probability $z_n' = 1 - z_n$ to m playing s_m'. After having constructed these beliefs, n will play a best response against $z_n s_m + z_n' s_m'$, in particular n will play the centroid of $B_n(z_n s_m + z_n' s_m')$, i.e. n chooses all best responses with equal probability. Denote this centroid by $b_n(z_n)$. Player m does not know n's beliefs z_n and, according to the principle of insufficient reason, m will assume that z_n is uniformly distributed on the interval $[0, 1]$. Hence, a priori m will expect n to play the strategy

$$p_n(s, s') = \int_0^1 b_n(z_n) dz_n, \tag{7.1}$$

and this prior is used as the starting point of the tracing procedure to determine the risk dominance relation between s and s'. Formally, s *risk dominates* s' if $T(p(s, s')) = s$. This risk dominance relation is reflexive and asymmetric. It need not be complete, nor transitive.

Write Ω^0 for the initial candidate set and define Ω^{t+1} as the set of all equilibria of Ω^t that are undominated within Ω^t. Let T be such that $\Omega^T = \Omega^{T+1}$. If Ω^T contains one element then that equilibrium is the solution of the game. Otherwise, there is a deadlock, which is resolved by a substitution step. Specifically, the equilibria in Ω^T are replaced by the trace of the centroid of Ω^T and the reduction process is continued with the new candidate set $\Omega^{T-1} \backslash \Omega^T \cup \{T(p(\Omega^T))\}$. In this way a decreasing sequence of candidates sets is obtained, hence, the process terminates with a unique equilibrium: The Harsanyi/Selten solution of the game.

7.1 Application to the Noisy Commitment Game

We have already seen in Section 5 that the Harsanyi/Selten theory need not select a noisy Stackelberg equilibrium since such an equilibrium need not belong to the initial candidate set. However, we also argued that this elimination step is not entirely convincing. Hence, the question remains whether the Stackelberg equilibrium can be eliminated by considerations of payoff dominance or risk dominance.

Proposition 2 implies that a noisy Stackelberg equilibrium, when unique[4], cannot be payoff dominated when the noise is small. Any Nash equilibrium outcome of the noisy game converges to a Nash outcome of the game in which the commitment is observed perfectly and among the latter the Stackelberg equilibrium is most preferred by player L. Consequently, it remains to address the question of whether the noisy Stackelberg equilibrium can be risk dominated. We have not been able to resolve the issue in its complete generality, however, for two important subclasses of games — 2×2 games and unanimity games — we can show that the (noisy) Stackelberg equilibrium risk dominates any other equilibrium of g^P when the noise $|P|$ is small. Hence, in these two classes of games, the modified Harsanyi/Selten theory (that does not restrict initial candidates to be primitive equilibria) selects the noisy Stackelberg equilibrium.

We say that $g =< I, J, u, v >$ is a unanimity game if (a) $I = J$, (b) $u_{ij} = v_{ij} = 0$ for all $i \neq j$, and (c) $u_{ii} > 0$ and $v_{ii} > 0$ for all i. We simplify notation by writing $u_i = u_{ii}$ and $v_i = v_{ii}$ and recall from (2.3) that $u_1 > u_i$ for $i \neq 1$. We also write "$\underline{i}$" for the strategy of player F in g^P that prescribes to respond to any signal $k \in I$ by playing $i \in I$.

Proposition 6 . *Let g be a unanimity game. Then the Stackelberg equilibrium (1,$\underline{1}$) risk dominates any other equilibrium of g^P when the noise P is small.*

Proof. We first show that (1,$\underline{1}$) risk dominates any other pure Nash equilibrium of g^P when $|P|$ is small. It suffices to show that (1,$\underline{1}$) risk dominates (2,$\underline{2}$). We first compute the bicentric prior that is used in the risk dominance comparison of these two equilibria. Let us first compute the prior p_L of player F. If F plays $z.\underline{1} + (1-z).\underline{2}$ then the best response of L is

$$b_L^P(z) = \begin{cases} 1 & \text{if } z > \quad u_2/(u_1+u_2) \\ 2 & \text{if } z < \quad u_2/(u_1+u_2) \end{cases} \tag{7.2}$$

hence, the prior of F is given by

$$p_L^P(i) = \begin{cases} u_1/(u_1+u_2) & \text{if } \quad i = 1 \\ u_2/(u_1+u_2) & \text{if } \quad i = 2 \end{cases} \tag{7.3}$$

Next we compute the prior of player L. If L plays $z.1 + (1-z).2$, then the best response of F depends on the message that F receives and on the size of the noise. However, since the posterior of F puts positive weight only on the actions 1 and 2 of player L, F will respond with either 1 or 2 at each possible message. Furthermore, if the noise is small, then F will respond to the message $i = 1$ (resp. $i = 2$) with the action 1 (resp. 2) for most values of z. Hence,without doing any computations, we may state that player L's prior p_F^P corresponds to a behavioral strategy f^0 of player F that is of the following form:

$$f_{ik}^0 = \begin{cases} \approx 1 & \text{if } \quad i = 1 \text{ and } k = 1 \\ \approx 1 & \text{if } \quad i = 2 \text{ and } k = 2 \\ 0 & \text{if } \quad k \notin \{1, 2\} \end{cases} \tag{7.4}$$

(f_{ik}^0 is the probability that F responds to signal i with action k.) Now, let the prior $p^P = (p_L^P, f^0)$ be as in (7.3), (7.4) and let the game g^{P,t,p^P} be as in (4.1). If t is sufficiently small, then the unique equilibrium $(s^{P,t}, f^{P,t})$ of this game is the best reply against the prior, hence

$$f_{ik}^{P,t} = \begin{cases} 1 & \text{if} \quad i = 1 \text{ and } k = 1 \\ 1 & \text{if} \quad i = 2 \text{ and } k = 2 \\ 0 & \text{if} \quad k \notin \{1, 2\} \end{cases} \tag{7.5}$$

and, provided that $|P|$ is sufficiently small,

$$s_1^{P,t} = 1. \tag{7.6}$$

We claim that, if we move along the distinguished curve in $\Gamma^P(p^P)$ by increasing t, then player F has to switch his strategy before player L does. The argument is simply that, if F does not switch from a strategy as in (7.5), then L is facing a convex combination of strategies of type (7.4) and (7.5), hence, this is just a strategy of type (7.4), against which the strategy from (7.6) is the unique best response. Hence, as t increases, player F's posterior beliefs put more and more weight on L playing "1" and gradually F switches to respond with "1" at more and more messages. Such changes in behavior of F however, do not necessitate a change in behavior of L: The strategy from (7.6) remains a best response. Consequently, if no equilibrium is reached yet, F will have to change again. Eventually (when t gets close to 1), F's posterior after the message "2" will put so much weight on L playing "1" that F will respond to that message by playing "1" as well. At that point in time we have obtained the equilibrium $(1,\underline{1})$ from g^P and no further adjustments are necessary. Hence, starting at the prior (7.3) – (7.4), the tracing procedure converges to $(1,\underline{1})$, so that $(1,\underline{1})$ risk dominates $(2,\underline{2})$. Consequently, the Stackelberg equilibrium risk dominates any pure equilibrium of g^P.

Next, let s' be a mixed strategy equilibrium of g^P. Proposition 2 implies that, if the noise is small, there exists an action $i \in I$ such that player L plays i with a probability very close to one. If $i = 1$, then $(1,\underline{1})$ is an equilibrium of g^{P,t,p^P} for all t. Hence, $(1,\underline{1})$ is the outcome of the tracing procedure. If $i \neq 1$, then the proof follows exactly the same line as above: Player L plays "1" for each value of t and player F switches several times until he finally responds to all messages by playing "1". □

Our final result is

Proposition 7 . *If g is 2×2 game and $|P|$ is small, then the noisy Stackelberg equilibrium of g^P risk dominates all other equilibria of this game.*

Proof. The result follows from Proposition 3 in case player F has a dominant strategy in g (g^P has only one equilibrium in this case). Hence, assume that F does not have a dominant strategy. Without loss of generality assume $b_1 = 1$ and $b_2 = 2$. In case g does not have any pure equilibria, the result again follows from Proposition 3 since g^P has a unique equilibrium in this case. There are three cases left to consider:

(i) (1,1) is the unique pure equilibrium of g.

(ii) (2,2) is the unique pure equilibrium of g.

(iii) both (1,1) and (2,2) are pure equilibria in g.

The first case is easy: It can be resolved by iterative elimination of strictly dominated strategies. (It should be obvious from the description of risk dominance that strategies that are iteratively strictly dominated cannot influence the risk dominance relationship.) The strategy "21" of player F (play $k \neq i$ in response to i for $i = 1, 2$) is strictly dominated and once this strategy has been eliminated, the strategy 1 becomes strictly dominant for player L. (Note that action 1 is dominant for L in g in case (i).) The third case is very much like the case considered in Proposition 6 and the proof proceeds along the same lines. We leave the details to the reader. In case (ii), g^P has three equilibria, viz. a mixed equilibrium with outcome close to (1,1), a mixed equilibrium with outcome close to (2,2), and the pure equilibrium (2,2). We have to show that the first equilibrium risk dominates the latter two. The proof follows from an argument as in Proposition 4. Namely, consider the bicentric prior p_F^P of player L in game g^P relevant for the comparison between the noisy Stackelberg equilibrium and the pure equilibrium (2,2). With p_F^P defined as in (7.1), we have that

$$\lim_{|P|\to 0} p_F^P(b) = 1, \tag{7.7}$$

since the strategy b of player F is a best response to the noisy Stackelberg equilibrium and is "almost" a best response to the pure equilibrium. Second, the bicentric prior of L relevant for the comparison between the noisy Stackelberg equilibrium and the mixed equilibrium is given by $p_F^P(b) = 1$, since the strategy b of player F is the unique best response against a strict convex combination of the two mixed equilibrium strategies of player L in g^P. In both cases, the remainder of the proof proceeds along similar lines as the proof of Proposition 4. □

Although we conjecture that the result from the Propositions 6 and 7 can be generalized to other classes of games, we have to admit that we have not been able to find a general proof. (We do not have a counterexample either.) However, we note that applying the tracing procedure can be rather complex, so that a multilateral procedure as that in Section 4 – in which the tracing procedure is applied only once – might be preferable to a theory in which one is forced to make a rather large number of bilateral comparisons. Furthermore, in order to apply the Harsanyi/Selten theory one has to first compute all (primitive) equilibria of the game. We were able to prove Proposition 4 without knowing this set of all equilibria.

8 Conclusion

From the fact that any pure Nash equilibrium of the noisy commitment game has to correspond to a Nash equilibrium of the underlying simultaneous move game (Proposition 1 in this paper), Kyle Bagwell concluded in his 1995 paper that slight noise eliminates any first mover advantages. In the concluding section of his paper, Bagwell writes

> "A central message of the paper is that the many predictions derived from models with commitment may require reconsideration. Apparently these

> predictions are valid only for settings in which the committed action is in fact *perfectly* observed by subsequent players. This requirement is quite stringent, and it would seem to be violated in a number of real-world settings to which popular commitment models are thought to apply" (Bagwell (1995, p. 278) emphasis in original).

While we agree with the observation that the assumption of perfect observability is stringent, we disagree with the statement that this assumption is crucial. In fact, we would claim that this paper shows that the assumption is largely inessential. Not only have we shown that the noisy game analyzed by Bagwell always has an equilibrium outcome that is close to the Stackelberg outcome of the game in which the commitment can be observed perfectly (Proposition 3), we have also given several arguments for why this outcome is the unique focal one, so that one would expect perfectly rational players to coordinate on it (Propositions 4, 6 and 7). Bagwell's claim relies essentially on the assumption that it is sufficient to look at pure strategy equilibria. As long as no convincing justification has been offered for this restriction, the present authors do not see any need to reconsider the fundamental game theoretic insight that the power to commit oneself may be beneficial.

References

BAGWELL, K. (1995). "Commitment and Observability in Games," *Games Econ. Behav.* **8**, 271-280.

BINMORE, K. (1987). "Modelling Rational Players," I and II. *Econ. and Philosophy* **3**, 179-214 and **4**, 9-55.

CHAKRAVORTI, B. AND Y. SPIEGEL (1993). "Commitment Under Imperfect Observability: Why it is better to have followers who know that they don't know rather than those who don't know that they don't know," Bellcore Economics DP 104.

VAN DAMME, E. (1987). *Stability and Perfection of Nash Equilibria.* Springer Verlag, Berlin. Second edition 1991.

VAN DAMME, E. (1995). "Strategic Equilibrium", forthcoming in: R.J. Aumann and S. Hart (eds), *Handbook in Game Theory,* Vol. III, North Holland Publ. Company.

HARSANYI, J. (1975). "The Tracing Procedure: A Bayesian approach to defining a solution for n-person noncooperative games," *Int. J. Game Theory* **4**, 61-94.

HARSANYI, J.C. (1995). "A New Look at Strategic Rationality: At Equilibrium Selection, Risk Dominance and Payoff Dominance," *Games Econ. Behav.* **8**, 91-122.

HARSANYI, J. AND R. SELTEN (1988). *A General Theory of Equilibrium Selection in Games.* Cambridge, Mass.: MIT Press.

HURKENS, S. (1994). "Learning by Forgetful Players: From primitive formations to persistent retracts," CentER DP 9437. Forthcoming in *Games Econ. Behav.*.

KALAI, E. AND D. SAMET (1984). "Persistent Equilibria in Strategic Games," *Int. J. Game Theory* **13**, 129-144.

KOHLBERG, E., AND J.-F. MERTENS (1986). "On the Strategic Stability of Equilibria", *Econometrica* **54**, 1003-1039.

MYERSON, R.B. (1978). "Refinements of the Nash Equilibrium Concept," *Int. J. Game Theory* **7**, 73-80.

NASH, J.F. (1950). *Non-Cooperative Games*, PhD Dissertation, Princeton University.

SCHANUEL, S.H., L.K. SIMON AND W.R. ZAME (1991). "The Algebraic Geometry of Games and the Tracing Procedure," in: R. Selten (ed.), *Game Equilibrium Models, Vol. 2: Methods, Morals and Markets*, Springer Verlag, 9-43.

SCHELLING, T. (1960). *The Strategy of Conflict.* Harvard University Press, Cambridge, Mass.

SELTEN, R. (1975). "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games", *Int. J. Game Theory* **4**, 25-55.

VON STACKELBERG, H. (1934). *Marktform und Gleichgewicht.* Springer Verlag, Vienna and Berlin.

Endnotes

1. Bagwell (1995) restricts himself to the case where player F has a unique best response to any pure action i of player L. He writes (on p. 275) that the basic results are most easily reported in this case, from which the reader might be tempted to conclude that his result (Proposition 1 in our paper) is also valid for games that do not satisfy this condition. That conclusion, however, is unwarranted as the following example shows

	W	E
T	2,4	0,4
B	3,0	1,2

Figure A1.

The unique pure equilibrium of the game g of Figure A1 is (B, E). However, (T, WE) is also a pure Nash equilibrium of the noisy commitment game and this equilibrium results in the outcome (T, W). The reader might object that the latter equilibrium is not credible since it is not perfect (although it certainly is sequential). This deficiency is easily eliminated by adding a third (dominated) strategy of player L to the game against which W is the unique best response of player F.

2. Harsanyi first eliminates dominated and duplicate strategies to ensure that all remaining pure strategies have stability sets with nonempty interiors, and that

intersections of interiors of different stability sets are empty. All definitions refer to such a reduced game.

3. Formally, $A' = A'_1 \times A'_2$ is a formation of a game γ if $B_n(s_m) \subset A'_n$ for all mixed strategies s_m of player m that assign positive probability only to $A'_m (m, n \in \{1, 2\})$. Minimal formations are called primitive. They exist, are disjoint and are finite in number.

4. In general a game may have more than one noisy Stackelberg equilibrium, i.e. the game g^P may have multiple mixed equilibria that all converge to the Stackelberg equilibrium when the noise vanishes. (An example is available from the authors upon request.) The games that we consider in this section do, however, have a unique noisy Stackelberg equilibrium.